STATISTIQUE AU LYCÉE

Volume 1 : Les outils de la statistique

Commission Inter-IREM
Statistique et Probabilités

Brochure APMEP n° 156
ISBN 2-912846-32-3

La présente brochure, due à la commission Inter-Irem « **Statistiques et Probabilités** », convie à des réflexions approfondies, dès le Collège et les Lycées, sur des sujets fondamentaux (outils descriptifs, modélisations et simulations…)

Sous l'efficace animation de Brigitte CHAPUT et Michel HENRY, il s'agit d'un travail collectif issu de confrontations de grande qualité.
La Commission a bien voulu en confier l'édition à l'APMEP, ainsi que celle, prochaine, du tome II (cf. p. 311).

Qu'elle en soit plus que chaleureusement remerciée, avec une mention particulière pour Brigitte CHAPUT et Michel HENRY et, par delà les auteurs, pour les IREM impliqués.

Je remercie tout aussi chaleureusement mon complice APMEP Jean Barbier qui a mis, avec son soin habituel, sa touche finale à une édition déjà fort bien préparée.

Henri BAREIL
responsable des Publications APMEP

STATISTIQUE
AU LYCÉE

VOLUME 1
LES OUTILS DE LA STATISTIQUE

Commission inter-IREM

Statistique et Probabilités

Coordination : Brigitte CHAPUT et Michel HENRY

Remerciements

Cette brochure de l'APMEP est une œuvre collective, fruit de trois années de travail de la Commission Inter-IREM *Statistique et Probabilités*. Ses membres ont contribué à sa réalisation par leurs remarques et par leurs interventions sur les différents sujets qui sont traités dans cet ouvrage. Outre les auteurs-rédacteurs des articles, d'autres collègues ont participé à leurs relectures et à leurs corrections, notamment : Bernadette DENYS (IREM de Paris 7), Jean-Pierre GRANGÉ (IREM de Besançon), Geneviève LORIDON (IREM de Dijon), Jean-Louis PIEDNOIR (Inspection Générale), Daniel VAGOST (IREM de Lorraine), Hervé VASSEUR (IREM d'Orléans-Tours). Que tous soient vivement remerciés pour leur participation assidue aux travaux de la Commission.

Ce livre a bénéficié de la bienveillance et de la patience des responsables des éditions de l'APMEP : Henri BAREIL et Christiane ZEHREN, qu'ils en soient chaleureusement remerciés.

Jean-Pierre RAOULT, Professeur des Universités émérite (Université de Marne-la-Vallée) et Président du Comité Scientifique des IREM, a bien voulu faire la dernière relecture des articles et nous faire bénéficier de ses conseils. Il nous a fait la gentillesse de les préfacer. L'ensemble de la Commission tient à lui exprimer ici sa reconnaissance amicale.

Sommaire

Première partie : Les outils de la description statistique

Deuxième partie : Simulations et modèles probabilistes

Annexes

Préface

Malgré la généralisation des cours de Statistique, en France et (souvent bien plus tôt que chez nous) dans d'autres pays, aux niveaux et dans les ordres d'enseignement les plus variés, en formations initiales ou en formations pour adultes, l'enseignement de cette discipline est réputé, à juste titre, pour être particulièrement délicat. Malgré l'abondance des études, colloques, rapports, il n'est, je pense, aucun segment de cet enseignement pour lequel on puisse affirmer avoir à proposer une solution idéale quant à l'imbrication du calcul des probabilités et de la statistique, quant à l'harmonisation de l'expérimental et du théorique ou encore quant à l'articulation de la statistique descriptive et de la statistique inférentielle. Pour ma part, en plus de quarante ans d'enseignement du Calcul des Probabilités et de la Statistique à différents publics, tous d'enseignement supérieur il est vrai (DEUG, Maîtrise ou DEA pour étudiants mathématiciens ou « mathématiciens appliqués », idem pour étudiants non mathématiciens -en biologie, sciences de l'environnement...-, IUT et IUP -formation d'informaticiens-, École d'Ingénieurs), j'ai pratiqué, voire échafaudé, de multiples maquettes sans connaître jamais une totale satisfaction.

Je suis donc tout à fait sensible aux inquiétudes, génératrices de réticences, de mes collègues de l'enseignement secondaire (collèges et, surtout, lycées) face aux différents avatars des programmes relatifs à ces branches des mathématiques et à leurs applications. Mon premier vœu en rédigeant cette préface est ainsi de « réconforter » le lecteur : s'il est souvent anxieux face à la manière de « faire passer » un enseignement qui est pour lui moins familier et moins balisé que celui du reste des programmes, qu'il sache que cette anxiété est naturelle. Mais cette compréhension n'a d'égale chez moi que ma conviction que l'apprentissage, par les jeunes des fondements de la réflexion et de la manipulation dans le domaine de l'aléatoire, est indispensable à la formation tant du travailleur que du citoyen. Il existe malgré tout un corpus de notions, de notations et de résultats qui font bien de la Statistique une science de type mathématique, et on y dispose d'une expérience pédagogique qui, même si elle est multiforme, peut être transmise utilement de collègue expérimenté à collègue plus débutant.

Un ouvrage tel que celui que présente ici la Commission Inter-IREM *Statistique et Probabilités* ne pouvait donc que susciter mon intérêt. Ce n'est pas la première fois que cette commission propose des recueils de ce type qui (c'est là le propre de l'action des IREM) se démarquent tant des manuels scolaires que des outils d'aide à l'enseignement proposés par les rédacteurs de programmes, par le fait que, à côté de rappels et compléments sur un certain nombre de notions du programme, ils fournissent à la fois des éclairages complémentaires (historiques, mathématiques) et des réflexions critiques sur ces programmes. Ce volume-ci, qui regroupe à la fois des contributions inédites et des articles (souvent réécrits pour l'occasion) déjà diffusés, par exemple dans *Repères IREM*, vient à son heure, au moment où les enseignants de mathématiques des lycées français ont eu à effectuer la mise en place, depuis la classe de seconde jusqu'à la terminale, de nouveaux programmes fondés sur une progression cohérente et innovante faisant démarrer la familiarisation avec l'aléatoire en Seconde par la pratique de la simulation.

La réflexion critique sur les programmes dans cet ouvrage me semble particulièrement riche et bienvenue, car sous-tendue à la fois par une connaissance intime des interrogations des enseignants et par une compréhension réelle de la logique des programmes actuels. Les lecteurs relèveront comme moi la forme interrogative qui est donnée aux articles qui comportent cette réflexion et qui ouvrent chacune des deux parties de l'ouvrage, consacrées respectivement à la DESCRIPTION STATISTIQUE et aux SIMULATIONS ET MODÈLES : d'une part *Pourquoi est-il si difficile d'enseigner la statistique ?* et d'autre part *Modélisation et simulation en classe, quel statut didactique ?* Un autre article de fond s'ouvre aussi sur un « pourquoi », même s'il s'agit ici de défendre avec énergie et pertinence, à mon avis, une position : *Pourquoi il ne faut pas laisser de côté les chapitres de statistique au collège*. On trouve aussi une analyse intéressante des difficultés didactiques au sein de l'article *Tests d'adéquation à une loi de probabilité, pratique des tests du Khi-deux*, où sont développés, sur un thème précis, deux des « malaises » principaux des enseignants de mathématiques français, face à la demande d'enseignement de la statistique à laquelle ils doivent répondre : où placer des « théorèmes » et comment les formuler ? Comment justifier la simulation dans la « vérification » d'une technique statistique ?

La qualité de nos enseignements ne peut qu'être améliorée par une perspective historique, même si sa traduction en termes pédagogiques n'est souvent que très partiellement réalisable. Cette conviction a présidé à de nombreux travaux popularisés par les IREM, elle est particulièrement justifiée s'agissant de la statistique, science jeune qui ne peut qu'être légitimée par la compréhension des questionnements qui l'ont suscitée et par la connaissance des maturations successives qui lui ont permis de répondre à ces questionnements. De plus certains des exemples concrets à l'origine des outils élaborés progressivement depuis environ un peu plus d'un siècle et aujourd'hui enseignés, restent simples et

parfaitement utilisables pédagogiquement, au prix éventuellement d'une certaine relecture. Relèvent ici de ce souci les articles *Expérimentation et simulation probabiliste* ou bien *Du modèle à sa réalisation. La planche de Galton réalise-t-elle vraiment une distribution binomiale ?* (article contenant un intéressant exemple d'affinement de modèle en y intégrant une réflexion de nature physique) ou encore *Théorie des erreurs, courbes en cloche et normalité.*

C'est aussi d'une préoccupation de « mise en perspective », au-delà de la lettre du programme, que relève *Quelques questions* à *propos des tables et générateurs aléatoires.* J'ai été pour ma part frappé, lors d'interventions auprès de collègues enseignant en lycées, de la force de leur désir de comprendre ce que fait véritablement un générateur de nombres aléatoires, désir dû, il me semble, à deux motifs complémentaires : d'une part la conviction (fondée) qu'il y a, « là-derrière » de belles mathématiques qu'il est bon de connaître, voire de communiquer à certains élèves, et d'autre part la gêne énorme devant le sentiment d'une forme de contradiction entre le caractère intrinsèquement déterministe de ces algorithmes et leur utilisation pour produire de l'aléatoire. La présence de cet article dans ce recueil est donc totalement justifiée et il me paraît apporter une masse d'informations et de commentaires précieux, même si je ne suis pas totalement d'accord avec sa conclusion, aux termes de laquelle il faut *surtout faire en sorte de tempérer (auprès des élèves)* « *l'effet boite noire* » *inhérent à l'utilisation d'une machine.* J'aurais tendance pour ma part à insister plutôt sur le fait que, comme par exemple pour un jet de dé, il y a dans l'élaboration du résultat lu toute une part de machinerie (qu'elle soit physique ou informatique) qui échappe à l'observateur et que c'est surtout la similitude des régularités produites par l'une ou par l'autre qui est à prendre en compte mathématiquement (aspect qui est lui aussi bien vu dans cet article).

Enfin on trouve dans ce recueil des articles plus directement liés à la lettre du programme, fournissant donc au lecteur, pour son enseignement, des éléments qu'il aurait certes souvent pu aller aussi glaner dans différents manuels d'enseignement supérieur (ou même, pour les aspects les plus élémentaires, secondaire) ou dans des descriptifs de logiciels statistiques, mais qui sont ici opportunément regroupés. Ces articles bénéficient souvent d'un accompagnement profitable par des exemples concrets ou des formulations d'exercices. Citons à cet égard, dans la première partie de l'ouvrage dévolue à la statistique descriptive, les deux articles qui se complètent mutuellement sur les observations unidimensionnelles, *Quartiles, déciles et tutti quantiles* et *Quelques pièges de la description d'une série statistique*, ainsi que, pour les observations multidimensionnelles, les deux textes *Description d'une série statistique à deux variables quantitatives : modélisation non probabiliste par les méthodes d'ajustement* et *Derrière la statistique, la géométrie*. Dans la deuxième partie, où se fait le lien entre le modèle probabiliste et la statistique, je citerai *Phénomènes gaussiens et loi normale*, de même que *Introduction aux tests*

d'hypothèses, exemples et enfin *Tests d'adéquation à une loi de probabilité, pratique des tests du Khi-deux.*

En conclusion, on ne peut que se réjouir du travail abondant et équilibré effectué par la Commission Inter-IREM *Statistique et Probabilités.* L'un de ses principaux mérites est d'être fondé sur le vécu des difficultés éprouvées par les enseignants de mathématiques des lycées français dans l'enseignement de la statistique ; j'ai la conviction qu'il doit aider nos collègues à affronter ces difficultés et j'espère, en tant que mathématicien-statisticien moi-même, qu'ils seront nombreux à en tirer même un véritable plaisir à communiquer à leurs élèves cette forme indispensable d'appréhension du réel.

Paris, mars 2005
Jean-Pierre Raoult
Président du Comité Scientifique des IREM

Présentation de l'ouvrage

Après *Autour de la modélisation en probabilités* et *Probabilités au lycée*, la Commission Inter-IREM *Statistique et Probabilités* a entrepris un travail de fond sur l'enseignement de la statistique tel qu'il est conçu dans le cadre des programmes des années 2000 des lycées. L'abondance des articles proposés par les membres de la commission, fruit de leurs pratiques dans leurs classes et de leurs recherches au sein de leurs IREM respectifs, nous a conduit à fractionner cette publication *Statistique au lycée* en deux volumes. Le présent volume est conçu comme une introduction, un débat et un élargissement autour des questions d'enseignement soulevées par les objectifs et la démarche adoptés dans l'ensemble des programmes de la seconde à la terminale. Le second volume, dont le sommaire figure en annexe, est plus centré sur les questions relatives à l'échantillonnage et aux situations de sondages, ainsi qu'à des exemples de simulations avec ou sans tableur.

Concernant la relation entre statistique et probabilités, les concepteurs des nouveaux programmes ont fait le choix d'une progression déterminée, allant de l'appréhension expérimentale des phénomènes aléatoires en seconde, notamment des fluctuations d'échantillonnage, à la modélisation probabiliste de situations simples en première, reposant sur une hypothèse d'équiprobabilité *quelque part*. La loi des grands nombres, livrée aux élèves sous une forme *vulgarisée*, est la clé de cette démarche reliant une expérience aléatoire à un modèle supposé adéquat pour la représenter. Elle permet d'avancer des conjectures informelles relatives à une population statistique à partir de données issues d'un échantillonnage, sans faire appel à des théorèmes de probabilités relevant de l'enseignement supérieur.

Le calcul de résumés statistiques permet de déterminer des valeurs suffisamment précises pour les paramètres intervenant dans les modèles probabilistes en jeu, et la simulation informatique conduit alors à la résolution expérimentale de certains problèmes d'inférences. L'introduction de la notion de loi de probabilité en première s'inscrit délibérément dans cette démarche de modélisation. C'est d'ailleurs l'objet d'une explication très claire donnée dans le document d'accompagnement du programme de première, rédigé par le GEPS à qui l'on peut rendre hommage pour cet important travail de qualité. Les exemples de lois continues proposés en terminale S peuvent ensuite être naturellement acceptés comme modèles théoriques pour des situations discrètes : la loi uniforme sur $[0\,,1]$ pour la génération de décimaux pseudo-aléatoires, la loi exponentielle comme modèle approché de la loi géométrique de l'attente de la première réussite dans un tirage répété d'une urne de Bernoulli, par exemple.

Le choix de cette progression, cohérente en elle-même, ne va pas sans poser de problèmes de nature didactique et épistémologique, comme l'a montré le vif débat qui a suivi la sortie du programme de seconde. La Commission Inter-IREM *Statistique et probabilités* a pris sa place dans ce débat. Tout en approuvant la démarche entreprise, nous avions souligné la difficulté didactique de priver les élèves de seconde du concept de probabilité pour rendre compte de leurs observations expérimentales. D'autant que de nombreuses recherches ont montré que cette notion s'installe naïvement chez les enfants qui, dans notre culture contemporaine, baignent dans l'aléatoire et son vocabulaire dès leur plus jeune âge. L'urne bicolore de Bernoulli est alors un générateur aléatoire paradigmatique, un objet didactique de référence, pour dégager assez tôt l'idée de probabilité, dès lors que l'initiation à la proportionnalité est engagée. Mais, pour nous, ce problème conjoncturel de progression tient au fait rédhibitoire d'un commencement trop tardif de l'éducation à l'aléatoire. Elle devrait s'installer dès le début du collège, voire avant[1].

Les remarques qui précèdent expliquent pour une part le contenu et l'organisation de ce premier volume en deux parties :

- Les outils de la description statistique.

- Simulations et modèles probabilistes.

On trouvera dans le deuxième volume quelques éléments sur l'estimation par intervalles de confiance et ses applications aux sondages ainsi que des exemples de traitements d'enquêtes.

Les références bibliographiques données dans les articles sont intégrées en fin d'ouvrage à une bibliographie relativement étoffée pour ce qui concerne les publications en français sur l'enseignement de la statistique. Deux index permettront de retrouver les pages où figurent les occurrences principales des termes généralement utilisés en statistique et en probabilités, ainsi que les noms des auteurs cités.

Nous espérons que cet ouvrage sera un outil efficace entre les mains de nos collègues qui investissent toutes leurs compétences et leur énergie pour que leurs élèves, de la seconde à la terminale, s'approprient des connaissances utiles en statistique et en probabilités.

Brigitte CHAPUT et Michel HENRY
Commission Inter-IREM
Statistique et Probabilités

[1] Cf. l'article de Bernard PARZYSZ : Peut-on envisager un enseignement de l'aléatoire au collège ? in *Le métier d'enseignant de mathématiques au tournant du XXI^e siècle.* Actes de l'université d'été de Marseille, APMEP, 2001.

Première partie :
les outils de la description statistique

Cette première partie s'ouvre avec deux articles de Jean Claude GIRARD abordant quelques grandes questions posées par cet enseignement de la statistique, accompagnées d'un plaidoyer pour que les concepts de base et le vocabulaire de la statistique descriptive soient bien installés à l'issue du collège.

Parmi ces concepts, ceux de médiane, quartiles et déciles, d'un usage de plus en plus répandu dans le domaine public, ne sont pas si simples à manipuler et nous avons pensé que le panorama des définitions en vogue présenté par Jean Claude GIRARD serait une aide pour les enseignants.

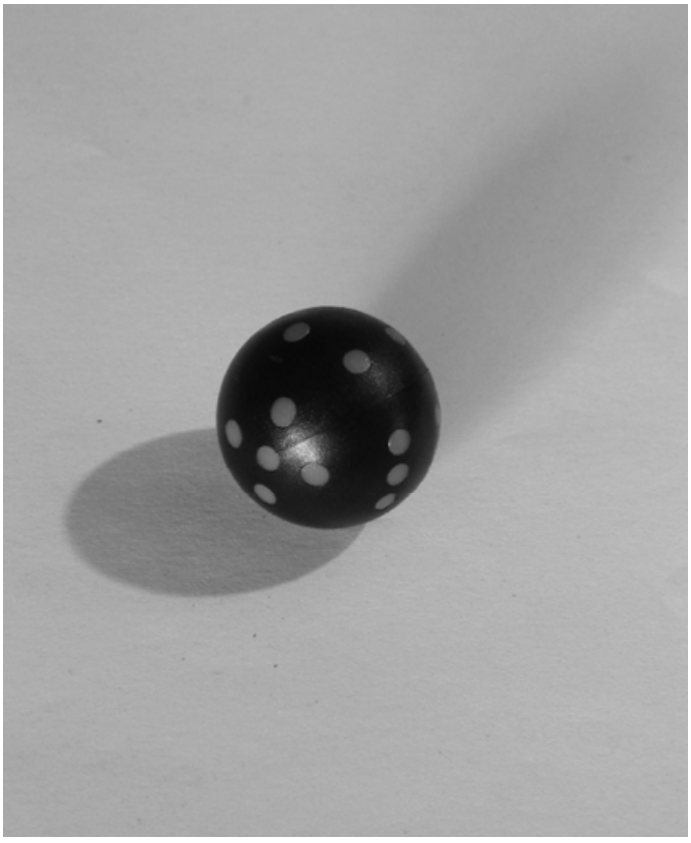
La description de séries statistiques à un caractère exploite divers outils et résumés : graphiques, histogrammes, tige et feuilles, paramètres de position et de dispersion, boîtes à pattes (ou à moustaches). Hubert RAYMONDAUD montre l'avantage de leur diversité, à condition d'éviter certains pièges. Il donne les modes d'emploi pour une bonne utilisation.

L'étude de données statistiques à caractère bidimensionnel est un objectif majeur de la série ES et de certains BTS. Stéphan MANGANELLI propose dans ce cadre un article illustré de nombreux exemples sur les questions de corrélation et d'ajustements… vers la régression, accompagné d'énoncés d'exercices et de devoirs.

On rencontre avec les séries chronologiques un aspect particulier de la description statistique. Brigitte CHAPUT en décrit quelques méthodes à partir de l'étude d'un exemple réel.

Enfin Brigitte CHAPUT et Jean Claude GIRARD présentent le cadre géométrique dans lequel opèrent les outils de la statistique exploratoire, appelée aussi analyse des données, mettant massivement à contribution l'algèbre linéaire et les structures euclidiennes dans $\mathbb{R}^n$.

Les méthodes modernes de la statistique descriptive, comme l'analyse en composantes principales ou l'analyse factorielle, ainsi que d'autres techniques mystérieuses que la plupart des collègues ont du mal à situer, ont été largement développées par les professionnels de la statistique, exploitant la puissance de calcul des ordinateurs. Nous limitant aux notions connexes aux programmes de l'enseignement secondaire, nous n'avons pas voulu aborder plus avant ces outils contemporains.



Pourquoi est-il si difficile d'enseigner la statistique ?

Jean Claude GIRARD

…aujourd'hui, il n'est pas exagéré de considérer la statistique en France comme une discipline émergeant difficilement[1].

Les derniers changements dans les programmes de lycée au cours de la période 2000-2002 ont marqué une révolution dans l'enseignement de la statistique. D'abord en terme de volume horaire (1/8 de l'année, en seconde, dit le programme officiel[2]), ensuite en raison des nouveautés présentées. Le programme de seconde introduit ainsi l'observation des fluctuations d'échantillonnage et la simulation informatique. En première les élèves étudient, pour la première fois au lycée, les quartiles et les diagrammes en boîtes. Les lois de probabilités y sont introduites par analogie avec les distributions de fréquence comme modélisation des situations aléatoires. En terminale, un test d'adéquation d'une distribution observée à une loi équirépartie s'inspirant du test du χ^2 est présenté. On peut donc voir dans ces changements une volonté de donner une importance beaucoup plus grande à l'observation statistique et de développer la liaison entre statistique et probabilités.

Ces changements soulèvent de nombreuses interrogations qui s'ajoutent aux questions que se posaient déjà les professeurs de mathématiques à propos de l'enseignement de la statistique. Parmi celles-ci, on peut citer : A quoi sert d'enseigner la statistique ? La statistique doit-elle être enseignée dans le cours de mathématique ? Comment enseigner ce qu'on n'a pas appris ? Comment mettre en place une démarche expérimentale en mathématiques ? Le professeur de mathématiques doit-il enseigner la modélisation ? Quel lien faire entre statistique et probabilité ? Comment définir, et enseigner, certains concepts comme le hasard ou la variabilité ? Comment concilier le raisonnement hypothético-déductif avec le raisonnement statistique ? Comment évaluer le travail des élèves en statistique ? Quelles conséquences didactiques tirer de toutes ces considérations pour l'enseignement de la statistique ? Que peut-on espérer d'un éventuel nouveau programme de collège ? etc.

---

[1] Académie des Sciences, *Rapport sur la science et la technologie n°8, juillet 2000.*

[2] Mathématiques, classe de seconde, nouveau programme applicable à compter de l'année scolaire 2000-2001, *B. O. n° 6, 12 août 1999, hors-série.*

La plupart de ces questions sont liées et on peut les regrouper en quatre points.

I - A quoi sert la statistique ? Faut-il l'enseigner dans le secondaire ?

La première question à se poser est effectivement celle-ci. Si on ne voit aucun intérêt à cette matière, toute autre discussion est sans objet.

Il est amusant de remarquer qu'on se pose plus rarement la question « A quoi sert l'Algèbre ? » ou « A quoi sert la Géométrie ? » ou encore « A quoi sert l'Histoire ? », « A quoi sert la Géographie ? ». La réponse est peut-être plus évidente dans ces cas mais il serait révélateur de comparer les réponses de chacun !

Si on met en perspective la vie professionnelle, la statistique est présente dans de nombreux domaines : scientifiques (physique, biologie, médecine...) mais également en psychologie, sociologie, sciences de l'éducation, sciences économiques, commerce, etc. Logiquement, cela explique que la statistique fasse partie des études correspondantes.

Comme l'affirme la Commission de Réflexion sur l'Enseignement des Mathématiques (CREM) :

« Les problématiques conduisant à des questions de nature statistique sont variées. La prise en compte de l'aléatoire a gagné presque tous les domaines : le contrôle de qualité en milieu industriel, la prévision des petits et des grands risques, l'élaboration de politiques de santé publique, les calculs financiers, etc. »[3].

Jean-Louis PIEDNOIR, Inspecteur Général de Mathématiques, ajoute avec humour, que les entreprises qui n'ont pas mis en place un contrôle statistique de la production[4] ont mis la clé sous la porte ! Ceci est une illustration de la *déraisonnable efficacité* des mathématiques, et de la statistique et des probabilités, en particulier.

Autant de raisons de les étudier. Pour caricaturer, de nombreuses filières d'enseignement supérieur en mathématiques (au moins pour devenir enseignant du secondaire et donc chargé de les enseigner) permettent d'y échapper !

L'enseignement de la statistique est une composante que l'on ne peut pas négliger de la formation du citoyen. L'utilisation abusive par les médias de moyennes et de pourcentages dont on ne sait pas sur quels référentiels ils ont été calculés, le bombardement incessant de résultats de sondages dont on ne sait ce qu'ils mesurent, de nombres dont on ne sait pas d'où ils proviennent, etc. imposent que l'élève, et ceci de façon assez précoce, apprenne le vocabulaire et les concepts de la statistique de façon à pouvoir comprendre et juger les arguments de ce type.

---

[3] *Rapport d'étape. Statistique et Probabilités*, 2002.

[4] SPC, Statistical Process Control, en anglais.

La statistique est aussi un langage que chacun doit maîtriser s'il ne veut pas être à l'écart du débat démocratique.

Il n'est pas nécessaire de chercher longtemps des exemples. La même radio, le même jour[5], annonce que le nombre de chômeurs a baissé tandis que le taux de chômage a augmenté, que la canicule a fait 11 435 morts et qu'un sondage fait apparaître que 60 % des Français sont pessimistes au sujet de l'avenir. On voit bien, dans chacun des cas, l'impression que l'on veut produire ou renforcer : le chômage augmente encore, la canicule a bien provoqué une catastrophe humanitaire, tout va mal en cette rentrée 2003. Le citoyen peut-il se laisser ainsi influencer par n'importe quel argument numérique sans s'interroger sur la façon dont ces chiffres ont été obtenus et sur leur signification réelle ? La statistique, et les mathématiques en général, ne peuvent servir de caution scientifique à n'importe quelle manipulation[6].

« Pour comprendre l'actualité, une formation à la statistique est aujourd'hui indispensable ; c'est une formation qui développe les qualités d'analyse et de synthèse et exerce le regard critique »[7].

CONDORCET (1743-1794), dont on oublie souvent qu'il était un mathématicien, croyait à l'amélioration du bien-être de tous par l'éducation et la science. Il voyait l'intérêt de la statistique qu'il appelait Mathématiques Sociales et il préconisait déjà son enseignement :

« Cette exposition montrera toute l'utilité de cette science ; on verra qu'aucun de nos intérêts individuels ou publics ne lui est étranger, qu'il n'en est aucun sur lequel elle nous donne des idées plus précises, des connaissances plus certaines ; on verra combien, si cette science était plus répandue, plus cultivée, elle contribuerait et au bonheur et au perfectionnement de l'espèce humaine.[8] *»*

L'enseignement de la statistique est aussi l'occasion de formation du raisonnement. Le raisonnement que l'on rencontre en statistique n'est pas du type hypothético-déductif (sauf dans les démonstrations de théorèmes, bien sûr, comme dans les autres parties des mathématiques). La situation générale en statistique est celle qui consiste à tirer des conclusions générales à partir de renseignements partiels ou, pour dire les choses autrement, d'inférer des conclusions sur une population à partir de résultats obtenus sur un échantillon extrait de cette

---

[5] France Info, le 30 août 2003.

[6] Un sommet semble avoir été atteint par un candidat à l'élection présidentielle de 2002 qui, voulant montrer le déclin de la France pendant le dernier septennat expliquait sans sourciller que notre pays était descendu du 4ème au 5ème rang mondial et du 3ème au 12ème rang en Europe sans se demander comment cela était possible !

[7] Commission de Réflexion sur l'Enseignement des Mathématiques, op. cit.

[8] *Elémens du calcul des probabilités et son application aux jeux de hasard, à la loterie et aux jugemens des hommes. Avec un discours sur les avantages des mathématiques sociales,* A Paris, chez Royez, libraire, An XIII.

population. Ceci est caractéristique du raisonnement inductif qui va du particulier au général. Ainsi, la *preuve* statistique n'est pas du même ordre que la preuve mathématique. On ne peut malheureusement qu'être sûr à 95 % (ou à 99 %…) mais rarement à 100 %. La plupart du temps on ne démontre pas qu'une hypothèse est vraie mais, plus modestement, on se contente de dire que l'on ne peut pas la rejeter ! Que penser d'une *Vérité* à 95 % ou d'une *Certitude* à 99 % ?

On conçoit que ce type de raisonnement puisse troubler particulièrement peut-être ceux qui sont bons en maths. Plutôt que de s'en émouvoir ou de le regretter, il vaudrait mieux se réjouir que soit donné aux élèves un autre type de raisonnement que l'on rencontre si souvent dans les autres sciences ou ailleurs.

Ce type de raisonnement (inductif) ne s'oppose pas d'ailleurs, mais complète et même utilise le raisonnement hypothético-déductif :

« *L'induction proprement dite suppose tout à la fois la connaissance des opérations déductives et celle du hasard lui-même. Le raisonnement inductif consistant précisément à trier ce qui est régulier et ce qui est fortuit pour organiser des régularités en un système de classe et de relations susceptibles d'un traitement déductif.*[9] »

Ne faut-il pas voir dans la difficulté de rentrer dans ce type de raisonnement, la conséquence d'un enseignement totalement déterministe que ce soit en maths, en physique ou même en économie ? Que penser de toutes les affirmations, preuves ou lois énoncées dans toutes les branches du savoir et à propos desquelles on peut se demander si leur Vérité est bien à 100 % ?

L'insuffisance de la formation des professeurs (en fonction depuis longtemps ou encore à l'IUFM) sur ce sujet est sans doute une explication de la reproduction de ces difficultés au fil des générations.

L'enseignement de la statistique est enfin l'occasion d'utiliser d'autres connaissances du champ mathématique et donc de les illustrer, de les développer, de leur donner du sens. On fait donc des mathématiques en dehors des calculs statistiques dans le cours de statistique[10]. Qu'on pense, par exemple, aux notions de proportionnalité, de pourcentage, d'échelle, de fonction, de représentation graphique, d'ensemble, etc. De plus, ces notions sont vues dans un autre cadre que celui où elles ont été introduites ou travaillées jusque là. Ceci ne peut que renforcer leur compréhension par les élèves. En conséquence, même si on accorde peu d'intérêt à la statistique, on ne perd pas complètement son temps en statistique.

---

9 PIAGET J., INHELDER B. *La genèse de l'idée de hasard chez l'enfant*, PUF, 1951.

10 Voir, par exemple, *Pourquoi il ne faut pas laisser de côté les chapitres de statistique au collège*, J.-C. GIRARD, Repères-IREM n°23, Avril 1996 réactualisé sous le même titre dans cette brochure.

II - La statistique doit-elle être enseignée dans le cours de mathématiques ?

Si la statistique a besoin de justifier son existence dans l'absolu, elle a également besoin de légitimation à l'intérieur des mathématiques. La première question que les élèves, et de nombreux professeurs également, se posent en effet, est : « Tout cela est-il réellement des mathématiques ? ». Cette interrogation n'est pas sans conséquences. Du côté des élèves, cela peut éloigner les meilleurs qui s'intéresseront aux choses plus nobles et du côté des professeurs cela peut conduire certains à limiter le temps imparti à la statistique à son strict minimum.

Ainsi, la présentation du cours ne prend pas l'aspect d'une suite de définitions précises et de théorèmes démontrés de façon rigoureuse et linéaire. Un des principes de base de la statistique, par exemple, est la variabilité des résultats dans la répétition d'une même épreuve aléatoire, basée sur des tirages *au hasard* dans une population ; mais qui peut donner une définition du *hasard* ? D'autres définitions ne sont pas très précises, contradictoires quelquefois d'un manuel à l'autre (comme celles des déciles ou des quartiles). Tout cela n'est donc pas très typique du cours de mathématiques. La conséquence en est la démobilisation des bons élèves en mathématiques concernant la partie statistique du cours.

De même tout ce qui est objet d'apprentissage en mathématiques est évalué de façon acceptable (ou au-moins accepté par la communauté scolaire) dans des exercices appropriés. Au contraire, l'évaluation en statistique n'est pas facile à construire entre l'exercice reproduisant exactement la situation d'apprentissage et le problème introduisant des difficultés imprévues. Comment éviter un contexte concret qui peut induire ses propres difficultés ? Comment évaluer, d'autre part, la compréhension de notions difficiles comme la fluctuation d'échantillonnage ou la loi des grands nombres ? Autant de questions qui peuvent renforcer les pratiques minimalistes c'est-à-dire l'apprentissage de techniques nécessaires et suffisantes pour faire les exercices du Baccalauréat.

Il ne faut pas nier la difficulté d'évaluation spécifique à l'enseignement de la statistique. Elle a fait l'objet d'une réflexion internationale qui a eu comme résultat un ouvrage entièrement consacré à ce sujet[11] (280 pages). La situation française ajoute une difficulté spécifique, à savoir la grande influence des sujets du Baccalauréat sur l'apprentissage au lycée (et même avant). Quelle sera l'influence d'un apprentissage expérimental sur la réussite aux exercices de statistique de l'examen si ceux-ci ne changent pas ?

---

[11] *Assessement Challenge in Statistics Education*, I. GAL & J.-B. GARFIELD (eds), International Statistical Institute (ISI) & International Association for Statistical Education (IASE), 1997.

III - Comment mettre en place une démarche expérimentale en mathématiques ?

Si on accepte le principe que la statistique fait bien partie du cours de mathématiques[12], il faut bien, si l'on veut donner du sens aux notions enseignées, les illustrer dans un contexte concret. Cela suppose que l'on sorte des mathématiques pour s'intéresser à une certaine réalité. Le problème qui se pose alors est celui de la modélisation c'est-à-dire de la pertinence de l'utilisation de tel ou tel concept dans telle ou telle réalité. Un problème fréquent est alors la confusion entre le modèle et la réalité[13].

Beaucoup de professeurs, là encore, pensent que ce n'est pas de leur ressort[14] et laissent leurs élèves aux prises avec l'implicite.

Le programme de seconde propose une démarche expérimentale utilisant les TICE (technologies d'information et de communication pour l'enseignement) dans chacun des trois chapitres de cette classe (statistique, calcul et fonction, géométrie). L'utilisation de l'outil informatique (tableur, logiciel de géométrie dynamique) ou des calculatrices graphiques « *multiplie… les possibilités d'expérimentation… Cet outil élargit les possibilités d'observation et de manipulation… Il donne la possibilité d'étudier une même notion sous une plus grande diversité d'aspect ; cela contribue à la démarche d'abstraction propre aux mathématiques et conduit à une meilleure compréhension.* »

Ceci est donc très nouveau et particulièrement déstabilisant pour les professeurs de mathématiques. Ce nouvel aspect de l'enseignement des mathématiques nécessitera d'être pris en compte dans la formation des professeurs. Mais il n'y a donc pas qu'en statistique qu'il convient de réfléchir à cette nouvelle approche qui s'apparente plus à l'enseignement des sciences appliquées qu'à celui des mathématiques.

Paradoxalement c'est peut-être en statistique que cela sera le plus facile à mettre en place. Comment lier statistique et probabilité, fréquence et probabilité, réalité et modèle, sans expérimentation sur une épreuve aléatoire (réelle ou simulée) ? Comment simuler une expérience un très grand nombre de fois sans ordinateur ou calculatrice ? La difficulté est alors d'ordre didactique : comment passer de l'expérience à la conceptualisation ?

---

[12] Pour s'en convaincre, il suffit de regarder les livres de Statistique d'un niveau supérieur dans lesquels on peut être amené à utiliser le modèle de géométrie euclidienne, d'autres distances plus exotiques, des raffinements de l'analyse ou de l'algèbre, etc…

[13] Voir par exemple, GIRARD J.-C. : Un exemple de confusion modèle-réalité, *Autour de la modélisation en probabilités*, Commission inter-IREM Statistique et Probabilité, coordination Michel HENRY, Collection Didactiques, Presses Universitaire Franc-Comtoises , 2001.

[14] Voir par exemple, GIRARD J.-C. : Le professeur de mathématiques doit-il enseigner la modélisation ?, *Repères-IREM n° 36*, p. 7-14, Topiques Editions, 1999.

Le déficit en formation des enseignants tant sur le contenu statistique que sur cette nouvelle approche expérimentale et plus généralement à propos de l'aléatoire est peut-être le frein le plus important à un enseignement efficace et généralisé de la statistique. En attendant que « *l'aléatoire fasse partie de la formation initiale des enseignants* »[15], « *la formation des enseignants de collège et lycée actuellement en poste est aujourd'hui un problème clé en ce qui concerne la formation citoyenne à l'aléatoire* »[16], « *cette formation est à créer* »[17].

IV - Quelles conséquences didactiques pour l'enseignement de la statistique ?

La variabilité est une notion clé de la statistique. Un des fondements de la méthode scientifique est de ne pas juger sur un seul exemple. Ni même à partir de deux ! L'éducation scientifique devrait permettre d'éviter ce genre de raisonnement (caricatural) : « j'ai eu des ennuis lundi dernier et le lundi précédent, j'en déduis que le lundi est un jour maudit ».

La variabilité n'est pas un concept *naturel* :

« *Bien que la variabilité dans le domaine du vivant soit, aux yeux de tous, une évidence …la force de cette évidence n'a d'égale que la facilité de l'oublier à chaque instant.*[18] »

Si on veut expliquer la variabilité des résultats d'une expérience répétée dans les mêmes conditions (germination de graines, fabrication d'un objet sur une même machine, etc.), on est bien obligé de sortir d'un modèle déterministe[19]. La Science a pour caractéristique de chercher des explications du Monde qui ne soient pas du domaine des croyances. La théorie des probabilités est un système possible d'explications construit sur l'idée de Hasard[20]. Sans qu'il soit possible de définir précisément ce terme[21], une construction intellectuelle permet d'obtenir certains résultats compatibles avec ce que l'on observe en réalité. Là encore, au lieu de s'en offusquer, n'y a-t-il pas urgence à former les élèves à ce type de modèle que l'on rencontre de plus en plus et dans tous les domaines : économie, médecine, contrôle de fabrication, etc.

---

[15] Commission de Réflexion sur l'Enseignement des Mathématiques, op. cit.
[16] ibid.
[17] ibid.
[18] SCHWARTZ D., *Le jeu de la science et du hasard*, Flammarion, 1994.
[19] C'est-à-dire un modèle dans lequel les mêmes conditions initiales dans une expérience conduisent au même résultat prévisible par le calcul (avec une certaine précision).
[20] Sur l'idée de Hasard, voir COURTEBRAS B. : Sur quelques conceptions du hasard, in *Autour de la modélisation en probabilités*, CII Statistique et probabilités, Besançon, PUFC, 2001.
[21] Mais donne-t-on une définition satisfaisante du "point" en géométrie ?

Cela suppose que les élèves se soient construit une certaine idée du hasard, en effet contrairement à une conception naïve :

« *Loin de s'imposer directement par le spectacle des faits expérimentaux, l'idée de* hasard *suppose une construction.*[22] »

Il serait donc nécessaire que les programmes prévoient une place et du temps pour cette construction :

« *L'expérience de cette dernière décennie montre que travailler sur l'aléatoire après le baccalauréat, sans une approche dans l'enseignement secondaire, est très difficile.*[23] »

On pourrait ajouter, sans doute, qu'il serait même souhaitable d'initier cette approche très tôt dans le secondaire[24]. Cela devrait inspirer les concepteurs de programmes pour ne pas renvoyer la première rencontre avec l'aléatoire en classe de seconde et au contraire prévoir des activités de ce type au collège[25]. La logique d'exposition voudrait que l'on fasse comprendre les notions par des activités et une présentation expérimentale (longtemps) avant de définir formellement les concepts. Ce n'est malheureusement pas le cas actuellement. Il ne faut pas chercher plus loin une des difficultés dans l'apprentissage des probabilités en première[26].

V - Conclusion

Nul doute que l'enseignement de la statistique soit difficile ! Sa spécificité fait qu'elle doit se battre pour sa légitimation à l'intérieur des mathématiques et à l'extérieur. Son évaluation est délicate. Elle souffre du manque de formation (et de motivation ?) des professeurs, elle traite de concepts difficiles comme le hasard ou la variabilité, se fonde sur un raisonnement qui lui est propre et débouche sur l'exploitation de modèles. Elle ne peut se satisfaire d'une présentation linéaire et dogmatique.

Au contraire, elle a tout à gagner de l'approche expérimentale proposée par les nouveaux programmes liée à l'activité de modélisation et basée sur la liaison statistique-probabilité. Cela nécessite toutefois une réflexion de fond sur la nature

---

[22] PIAGET J., INHELDER B. *ibid.*

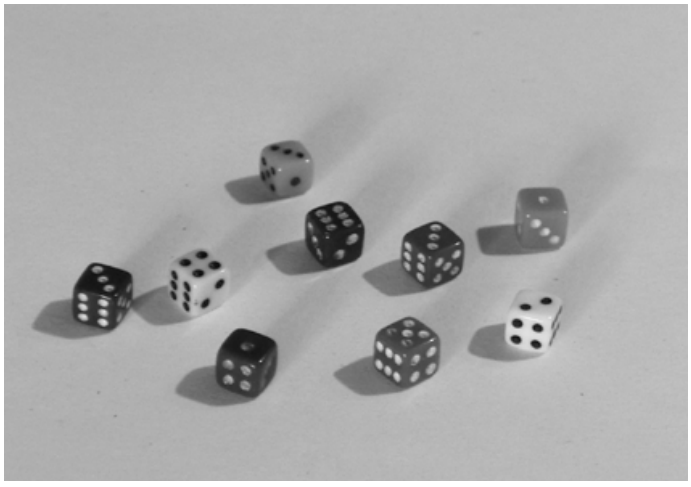
[23] Commission de Réflexion sur l'Enseignement des Mathématiques, op. cit.

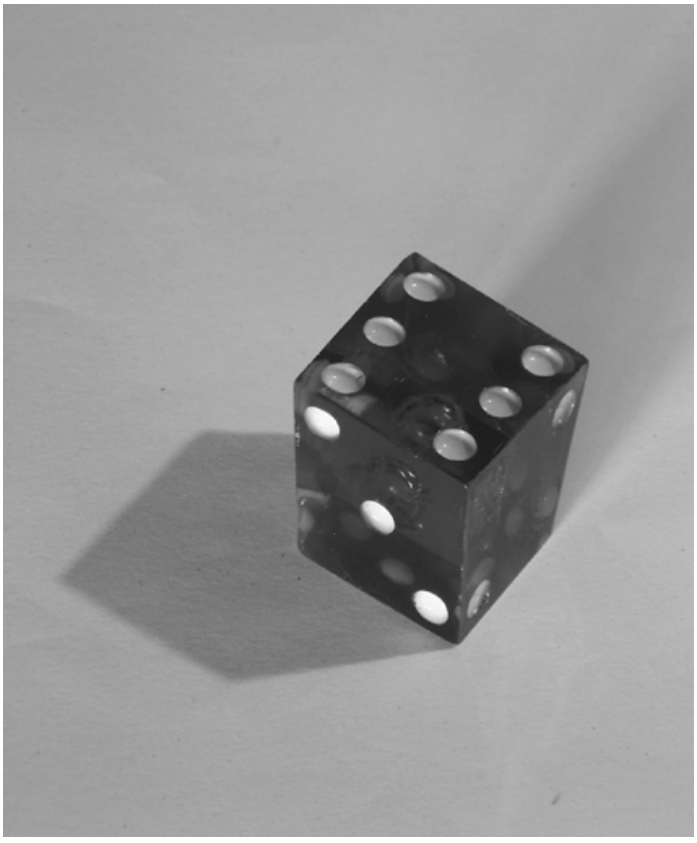
[24] Voir par exemple, GIRARD J.-C., HENRY M., PARSYSZ B., PICHARD J.-F. : Quelle place pour l'aléatoire au collège ?, *Repères-IREM n° 42*, p. 27-43, Topiques Editions, 2001.

[25] Une phrase du dernier programme du cycle 3 de l'école primaire (2001-2003) permet d'espérer sur ce point : « *Quelques exemples de phénomènes aléatoires peuvent être proposés dans la perspective de faire apparaître des régularités (par exemple, lancers d'une pièce ou d'un dé, lancers de deux dés dont on fait la somme)* ».

[26] Voir par exemple, GIRARD J.-C. : Difficultés et obstacles dans l'enseignement des probabilités, *Probabilités au lycée*, Commission Inter-IREM Statistique et Probabilité, coordination Brigitte CHAPUT, Brochure APMEP n°143, 2003.

de la modélisation, du statut de l'expérience et de la simulation et sur la difficulté pour l'élève dans le passage de l'observation à la théorisation. C'est, nous semble-t-il, l'enjeu de la décennie à venir. La commission Inter-IREM *Statistique et Probabilités* a conçu cette brochure comme un outil permettant aux professeurs de mathématiques de surmonter les difficultés évoquées. Les autres articles qui la composent suggèrent des réponses aux questions posées dans celui-ci, pour contribuer à la réussite de cet enseignement de la statistique et en fin de compte pour apporter aux élèves une formation en prise avec le monde contemporain.





Pourquoi il ne faut pas laisser de côté les chapitres de statistique au collège[1]

Jean Claude GIRARD

L'idée de cet article part d'un constat : les chapitres de statistique au collège sont souvent négligés, reportés à la fin de l'année ou tout simplement *sautés* sous prétexte que l'on n'a pas le temps de tout faire ! L'étude sérieuse en est alors différée d'année en année jusqu'à ce qu'on considère (en seconde généralement) que tout a été vu auparavant ! On observe d'ailleurs la même attitude pour l'utilisation de la calculatrice, dont on peut trouver l'idée très intéressante et pourtant reporter l'utilisation chaque année à la suivante par manque de temps ou parce que c'est trop tôt ! A cet égard, la statistique a rejoint la géométrie dans l'espace, fréquemment repoussée le plus loin possible dans l'année, rapidement traitée, éventuellement pas traitée du tout suivant le temps disponible.

La première raison de ce choix (parce que c'est un choix !) est que beaucoup de professeurs se sentent moins à l'aise dans ce chapitre, *moins mathématique*, que dans les autres, mais cela ne me paraît pas être la raison principale. Tout professeur consciencieux oublierait, en effet, ses états d'âme s'il était convaincu de l'intérêt de cette partie du programme et des difficultés qu'elle présente pour les élèves. Ce n'est malheureusement pas le cas.

Je vois au moins trois intérêts majeurs à développer l'enseignement de la statistique en tout cas pour qu'il atteigne le niveau que le programme lui assigne :

- au niveau des graphiques, en liaison avec différentes parties du programme de mathématiques et pas seulement pour servir d'outil à d'autres matières car, « *L'enseignement des statistiques contribue au* développement des compétences en mathématiques » (Document d'accompagnement du programme de troisième).
- au niveau des calculs (fréquences, moyenne, médiane) en liaison avec l'*idée de distribution statistique*,
- au niveau conceptuel en liaison avec l'*idée de hasard et de variabilité* des résultats dans une expérience (que l'on qualifiera alors d'aléatoire).

[1] Cette nouvelle version d'un article paru sous le même titre dans le n° 23 de la revue Repères-IREM (avril 1996) tient compte des changements de programme en lycée (septembre 2000).

L'étude de ces trois aspects de la statistique peut concourir au développement intellectuel des élèves et en particulier à l'aspect *formation du citoyen* confronté de plus en plus aux statistiques (graphiques, pourcentages, moyennes, sondages, etc.). L'objectif visé serait que les élèves se posent eux-mêmes des questions sur ce qu'ils voient ou entendent (chiffres ou graphiques). Cette étude me semble également indispensable en vue de faciliter l'*enseignement des probabilités* en première et terminale, si l'on ne veut pas se contenter de constater à ce moment là « qu'ils ont des difficultés ».

I - Les graphiques

C'est la partie de la statistique qui est la moins souvent *oubliée* car elle a des applications dans les autres matières et, de plus, elle fait assez souvent l'objet de questions au Brevet des collèges. D'autre part, on saisit l'occasion de la construction des graphiques statistiques (camemberts, barres, histogrammes) pour réinvestir la notion de proportionnalité sous ses différentes formes : pourcentages, échelles, règle de trois. L'hypothèse implicite est que ces graphiques ne posent pas de problèmes (autres que ceux liés à la proportionnalité) aux élèves. Et pourtant, en dehors des difficultés purement statistiques (définition des variables, récoltes des données), il reste beaucoup de points d'interrogation.

D'abord sur le sens des graphiques eux-mêmes :

- Quel est l'avantage d'un graphique sur un tableau de valeurs ?
- Le graphique sert-il d'illustration ou permet-il de découvrir une structure des données que le tableau ne mettait pas en évidence ?
- Peut-on repasser du graphique au tableau ?
- Quelle perception de la réalité a-t-on en regardant un graphique ?
- Pourquoi tel graphique plutôt qu'un autre ? Dans quels cas, chacun est-il pertinent ?

Ensuite sur d'autres notions qui renvoient à différents domaines mathématiques :

- Les camemberts utilisent les notions d'angle et de mesure d'angle qui ne sont pas toujours acquises. Comment peut-on prendre en compte cet état de fait ? Que représente le disque complet ? Autrement dit, quel est l'ensemble sur lequel on calcule les pourcentages ?
- Les histogrammes et les graphiques en barres ou en bâtons utilisent une échelle verticale sur laquelle on porte des effectifs ou des fréquences. Sur quel ensemble de référence ces fréquences ont-elles été calculées ?

– Lorsque l'on représente des *variations*, sont-elles calculées de façon absolue ou relativement à une valeur de référence ?

Exemple (extrait d'un livre de CM1 : Objectif Calcul - Hatier)

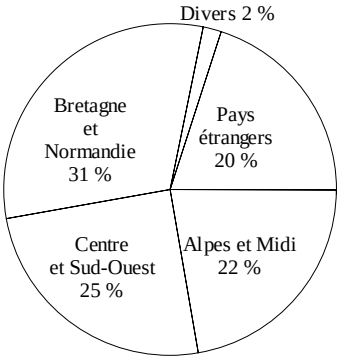
Le livre pose les questions suivantes :

1) *Observe ce graphique*

2) *Essaie de le lire*

3) *Quels renseignements donne-t-il ?*

4) *Essaie de traduire ce graphique par un tableau de nombres.*



Les vacances des français

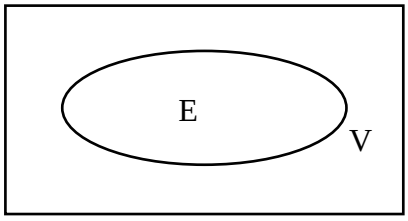
On pourrait aussi demander (en CM2, en 6ème ou plus tard !) :

– *Sur quoi sont calculés les pourcentages ?*

– *Est-ce 20 % des français qui partent en vacances à l'étranger ou 20 % de ceux qui partent en vacances qui vont à l'étranger ?*

– *Peut-on calculer combien de français partent à l'étranger ? Combien partent en vacances ?, etc.*

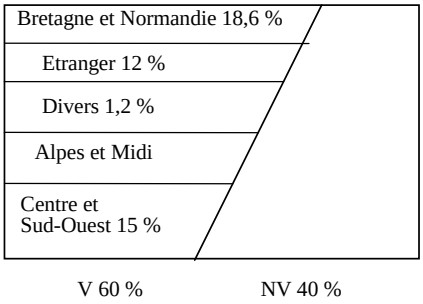
Cela pourrait être l'occasion d'une initiation aux représentations ensemblistes :



On peut raisonner sur la population française ou, pour simplifier, sur 100 personnes. Si l'on considère que 60 % des français partent en vacances, les 20 % qui vont à l'étranger représentent en fait 20 % de 60 %, c'est-à-dire 12 % de la population. La question fondamentale est : calcule-t-on les pourcentages sur l'ensemble de la population ou sur l'ensemble des français qui partent en vacances ? Ce genre de questions permet de donner du sens aux pourcentages bien plus que l'entraînement à la virtuosité dans les calculs.

Ces questions sont une préparation à l'étude des probabilités car on retrouvera les mêmes problèmes lorsque l'on raisonnera (en première) en termes de probabilités : un français étant choisi au hasard, quelle est la probabilité qu'il prenne ses vacances à l'étranger si l'on sait qu'il part en vacances ? (probabilité

conditionnelle de E sachant V, soit 20 %) ou quelle est la probabilité pour le même français de partir en vacances à l'étranger (E et V, soit 12 %) ?



D'ailleurs de nombreux problèmes de probabilité sur les ensembles finis se ramènent à des problèmes de fréquences ou de pourcentages.

Exemple (extrait du livre de terminale ES Déclic, collection Hachette, 1994)

Lors d'un sondage auprès de 24 000 personnes, 14 280 sont parties en vacances et 5 340 sont parties en vacances d'hiver.

Calculer la probabilité des événements suivants :

a) *« une personne, prise au hasard, est partie en vacances » ;*

b) *« une personne, prise au hasard, est partie en vacances d'hiver » ;*

c) *« une personne, partie en vacances, est partie en vacances d'hiver ».*

On peut remarquer que les probabilités présentent les mêmes difficultés que les pourcentages au niveau de l'ensemble de référence.

On peut les ajouter (ou les soustraire) si les calculs ont été faits sur les mêmes ensembles de référence : $P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ et } B)$.

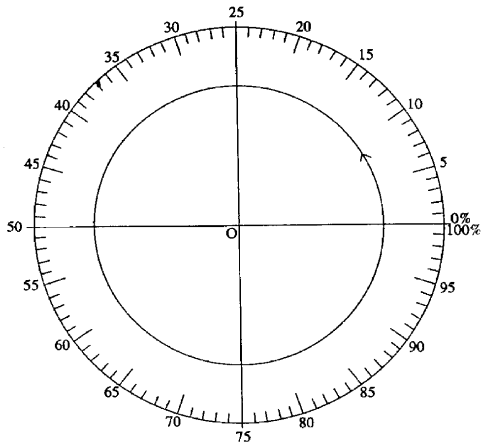
On les multiplie si un calcul a été fait sur un premier ensemble et l'autre sur un sous-ensemble de celui-ci : $P(A \text{ et } B) = P_B(A) \times P(B)$.

Ces questions concourent également à l'apprentissage de la lecture de graphiques ; celle-ci est au moins aussi importante que la construction. A quoi peut-il servir de construire des graphiques si l'on ne sait pas lire les graphiques déjà construits ? Quelle idée un élève se fait-il en regardant un histogramme ou un camembert ? L'a-t-on entraîné à lire un graphique ? A-t-il une perception globale des quantités représentées ou se fait-il une idée des unes par rapport aux autres ? ou par rapport à un tout ? On peut faire le pari que la lecture d'un graphique statistique est du même ordre que la lecture d'une figure de géométrie dans l'espace. Le décryptage n'est pas inné. La première perception est visuelle mais l'interprétation est cognitive, elle demande des connaissances. La lecture de l'expert n'est pas celle

de l'élève[2]. Il doit donc y avoir apprentissage de la lecture d'un graphique statistique. Les conceptions d'un élève sont souvent dans la comparaison plus grand, plus petit, et ceci sur des grandeurs prises dans l'absolu. L'objectif devrait être de les amener à comparer les valeurs les unes par rapport aux autres ou par rapport à un tout, c'est-à-dire à raisonner en valeur relative, en pourcentage, et alors, être capable d'identifier l'ensemble de référence ?

Par conséquent, il pourrait y avoir un grand intérêt à travailler les graphiques statistiques autrement que comme application de la proportionnalité. Ils devraient être un moyen de développer ce concept lui-même, les deux concepts s'éclairant mutuellement.

Tout comme la notion d'angle ne saurait être acquise sans en avoir une bonne image mentale, il me semble nécessaire de faire acquérir une image mentale d'un pourcentage. Cela nécessite un apprentissage. Des séquences peuvent être construites[3] à partir de la lecture et de la construction de graphiques statistiques en utilisant par exemple un rapporteur à pourcentage[4].



La perception de la proportionnalité n'est pas la même sur les longueurs que sur les aires[5]. Pour ceux qui sont plus sensibles à une *vision* linéaire de la proportionnalité, on peut aussi travailler sur les barres.

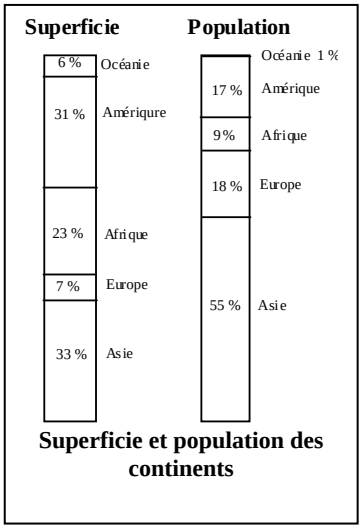
---

[2] Voir à ce sujet l'article de Jacques COURIVAUD, *Le traitement graphique des images de géométrie*, Repères-IREM n° 4, juillet 1991.

[3] Voir, par exemple, l'article de Daniel GROS, *Une enquête statistique au service de la proportionnalité*, Repères-IREM *n°44*, juillet 2001.

[4] Matériel en vente à l'IREM de LYON.

[5] Et encore moins sur les graphiques en perspectives, qui sont la plupart du temps faux d'un point de vue mathématique. On lira avec profit l'article de Gérard PORNIN *Des impôts à l'ellipse* dans *Des chiffres et des lettres au collège*, bulletin Inter-IREM Premier Cycle 1991-1992, dans lequel on présente une activité statistique dont les objectifs sont géométriques (théorème de Thalès, trigonométrie, cercle circonscrit, angles, symétries, tracés).



Superficie et population des continents

Exemple : (Objectif Calcul CM1)

– *Que représente la longueur de chaque barre ?*

– *Sur quoi ont été calculés les pourcentages ?*

– *Peut-on comparer ces différents pourcentages ?*

– *Quelle idée veut donner ce graphique ?*

II - Les paramètres et les distributions statistiques

On étudie classiquement au collège les effectifs et les fréquences d'apparition des différentes modalités d'un caractère qualitatif ou des valeurs prises par un caractère quantitatif. Dans ce dernier cas on calcul la moyenne (pondérée). La médiane est au programme de troisième mais son étude est souvent esquivée, pour plusieurs raisons : son calcul est moins automatique et beaucoup de professeurs ne voient pas l'utilité de ce paramètre, qui pourtant a l'intérêt de ne dépendre que de l'ordre total des données et non de leurs valeurs relatives.

On peut se demander, en effet, pourquoi calculer certains paramètres d'une série statistique ? Si l'on s'en tient au calcul de la moyenne, par exemple, il faut reconnaître que cela n'a pas beaucoup de sens, et même dans certains cas, pas du tout. L'objectif est de donner une idée d'une série statistique par une valeur numérique ou de comparer deux séries statistiques. On ne peut le faire avec les seules moyennes. Que peut-on dire, par exemple, d'un endroit où la température annuelle moyenne est de 20° ? La sensation ne sera pas du tout la même si les températures sont situées toute l'année entre 18° et 22° ou si elles évoluent entre – 40° l'hiver et + 30° l'été.

La moyenne ne prend son sens que si elle est associée à une mesure de la dispersion des valeurs de la série. Par exemple, l'écart-type qui prend en compte les écarts (par rapport à la moyenne) de chacune des valeurs de la série. L'inconvénient de ce paramètre est qu'il n'a pas de représentation concrète simple, qu'il est long à calculer, qu'il ne figure qu'au programme de première et, de plus, étant lié à la moyenne il est sensible, comme cette dernière, à des valeurs

anormalement grandes ou petites. Il existe heureusement d'autres paramètres de dispersion. Lorsque l'on porte dans un bulletin scolaire la moyenne, la note la plus basse et la note la plus haute, on caractérise la distribution des notes par un paramètre de tendance centrale, sa moyenne, et par un paramètre de dispersion, son étendue, c'est-à-dire l'écart entre le minimum et le maximum de la série. Ce dernier paramètre est simple à comprendre, d'un calcul aisé et est au programme de troisième ! L'inconvénient, cette fois, est qu'il est assez frustre et encore plus sensible aux valeurs extrêmes.

Une alternative, réalisable en collège, est de caractériser une série statistique par sa médiane, pour la tendance centrale, et par l'étendue de la *moitié centrale* pour la dispersion. Le programme de Troisième précise en effet : (Pour la notion de dispersion) *Le programme se limite à l'étendue d'une série statistique ou d'une partie de celle-ci*. Les valeurs de la série de départ étant rangées dans l'ordre croissant, il suffit alors de calculer la valeur médiane Me puis les médianes M_1 et M_2 des deux sous-séries qu'elle détermine[6]. L'étendue $M_2 - M_1$ mesure la dispersion de la série initiale d'une façon moins sensible aux valeurs extrêmes[7]. On a finalement une notion assez simple, de calcul relativement aisé et qui présente, de plus, le double avantage de réinvestir la médiane et de se prêter à une représentation graphique (voir plus loin). La conjonction de ces paramètres permet alors d'analyser une série statistique ainsi que de comparer des séries statistiques d'un double point de vue (tendance centrale et dispersion) en donnant du sens à ces deux concepts.

Exemple : D'après les statistiques de l'éducation nationale[8], le nombre d'élèves par section de sixième (établissements publics de France métropolitaine en 1989-1990), se répartit comme suit :

18 et –	19 à 23	24	25	26 à 27	28 à 29	30 et +
4,9 %	24 %	14,8 %	14,6 %	25,1 %	12,9 %	3,7 %

La moyenne (même source) s'élève à 24,6 élèves par section.

On peut se livrer dans un premier temps à des calculs classiques sur les pourcentages et les moyennes.

1° - Combien de sections de sixième avec 24 élèves ?, 25 élèves ?, etc. (Il faut donc transformer les pourcentages en effectifs en admettant par exemple que le nombre total de sections de sixième est de 30 000.)

---

[6] Lorsque plusieurs termes de la série statistique sont égaux à Me, on considère celui qui partage la série ordonnée en deux sous-série d'égal effectif, de même pour M_1 et M_2.

[7] Ce concept sera précisé en première à partir du calcul des quartiles.

[8] *Repères et références statistiques sur les enseignements et la formation 1991-1992*, Ministère de l'Education Nationale, Direction de l'Evaluation et de la Prospective, 1993.

2° - Comment calculer la moyenne lorsque les données sont regroupées en classes (de nombres d'élèves par section) et, qui plus est, que les classes extrêmes ne sont pas bornées ?

(On prend habituellement comme valeur de chaque classe, le centre de la classe en convenant, par exemple, qu'il n'y a pas de section comportant moins de 16 élèves, ni plus de 32, ce qui donne comme valeurs des centres de classes : 17 ; 21 ; 24 ; 25 ; 26,5 ; 28,5 ; 31 et comme moyenne 24,55.)

On peut remarquer que ceci n'est qu'une valeur approchée puisque l'on a perdu des informations en regroupant les données alors que l'on peut penser que la valeur du ministère a été calculée à partir des données brutes et qu'elle est exacte (mais arrondie !).

3° - Comme on l'a déjà fait remarquer, la moyenne ne donne pas de renseignements sur les variations du nombre d'élèves par section. On peut passer alors à l'analyse de la série par les paramètres proposés au début de ce paragraphe.

Si l'on considère que la série comporte 30 000 sections de sixième, alors la médiane est le nombre d'élèves de la $15\,000^{ème}$ section de la série ordonnée (en principe la moyenne entre la $15\,000^{ème}$ et la $15\,001^{ème}$!). Il convient de ne pas confondre (erreur fréquente chez les élèves) les rangs des données (dans un classement dans l'ordre croissant par exemple) et les valeurs de ces données.

Pour préciser notre analyse, faisons l'hypothèse d'équirépartition supposant que dans les classes regroupant plusieurs nombres possibles d'élèves par section (26 à 27 par exemple), il y a autant de sections pour chacun des effectifs d'élèves considérés (hypothèse douteuse pour les classes extrêmes, mais sans grandes conséquences pour la suite). La série ordonnée des effectifs des élèves des 30 000 sections se présente alors ainsi :

numéros d'ordre	1	2	.....	15 000	15 001	.....	29 999	30 000
valeurs des données	16	16	.....	25	25	.....	32	32

La série est alors partagée en deux sous-séries des effectifs de 15 000 sections que l'on peut de nouveau partager en deux par leurs médianes respectives :

numéros d'ordre	1	2	.....	7 500	7 501	.....	15 000
valeurs des données	16	16	.....	23	23	.....	25

numéros d'ordre	15 001	15 002	.....	22 502	22 501	.....	30 000
valeurs des données	25	25	.....	27	27	.....	32

La médiane M_1 de la première série est 23, celle de la deuxième série M_2 est 27.

Ces deux valeurs associées à la médiane Me de la série d'origine, partagent cette série en quatre sous-séries de même taille :

numéros d'ordre	1	.....	7 500	.....	15 000	.....	22 500	.....	30 000
valeurs des données	16	.....	23	.....	25	.....	27	.....	32
% des valeurs	25 %			25 %			25 %		

L'étendue de la moitié centrale de 15 000 sections est alors $M_2 - M_1 = 4$.

Avec notre hypothèse d'uniformité, on peut retrouver ces trois valeurs en utilisant le tableau des fréquences cumulées.

nombre d'élèves	18 et –	19	20	21	22	23	24
fréquence des sections	4,9 %	4,8 %	4,8 %	4,8 %	4,8 %	4,8 %	14,8 %
fréquence cumulée croissante	4,9 %	9,7 %	14,5 %	19,3 %	24,1 %	28,9 %	43,7 %

nombre d'élèves	25	26	27	28	29	30 et +
fréquence des sections	14,6 %	12,55 %	12,55 %	6,45 %	6,45 %	3,7 %
fréquence cumulée croissante	58,3 %	70,85 %	83,4 %	89,85 %	96,3 %	100 %

Les 50 % sont atteints pour une section d'effectif « 25 élèves », donc Me = 25.

Les 25 % sont atteints pour une section d'effectif « 23 élèves », donc $M_1 = 23$.

Les 75 % sont atteints pour une section d'effectif « 27 élèves », donc $M_2 = 27$.

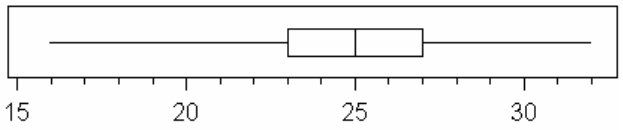
On remarque que 50 % des valeurs d'effectifs, correspondant à la *moitié centrale* de la série, sont dans l'intervalle [23 ; 27].

On peut représenter la série par un graphique, très utilisé dans les pays anglo-saxons mais encore peu répandu en France[9], mais appelé à le devenir puisque figurant au programme de première depuis septembre 2001, et appelé box-plot[10] ou graphique en boîte ou encore boîte à moustaches, qui donne une illustration de la tendance centrale (par la médiane : le trait à l'intérieur de la boîte) et de la dispersion de la série (par l'étendue totale : la longueur totale entre les extrémités des moustaches soit 32 – 16, et l'étendue de la *moitié centrale* : la longueur de la boîte soit 27 – 23)[11].

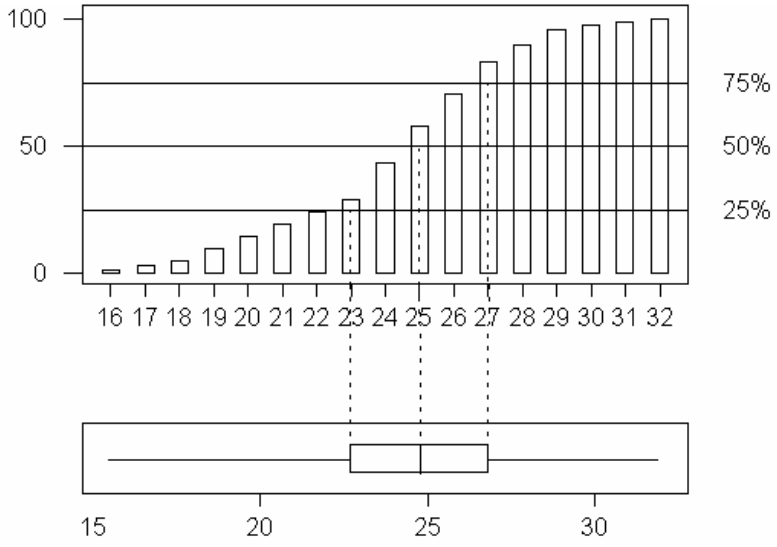
[9] bien que faisant partie des potentialités de certaines calculatrices comme les TI 83, Casio Graph 30 ou HP 38G

[10] TUKEY, J. W. : *Exploratory Data Analysis*, Addison Wesley, Reading, MA, 1977.

[11] Pour plus d'explications sur la construction des graphiques en boîte, voir par exemple, GIRARD, J. C. : La médiane, pour quoi faire ? Un exemple d'utilisation : les boîtes de dispersion, *Enseigner la statistique du CM à la seconde. Pourquoi ? Comment ?*, IREM de Lyon , 1998. Voir aussi dans ce même volume, l'article : Quelques pièges de la description d'une série statistique.

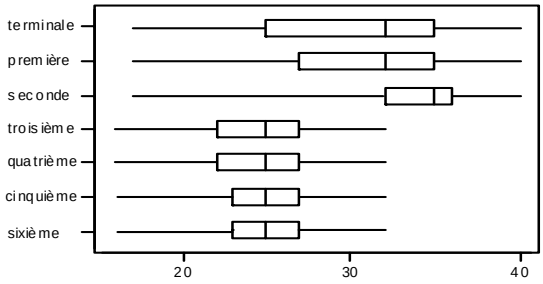


Le passage des fréquences cumulées au graphique en boîte peut se faire sur le graphique suivant.



On peut alors voir facilement (ce qu'on peut lire dans les données initiales, mais il faut avoir l'idée d'aller le chercher, que plus de 50 % des classes de sixième ont entre 23 et 27 élèves.

Ce genre de graphique prend tout son intérêt lorsque l'on veut comparer plusieurs séries statistiques. Par exemple si l'on veut analyser les effectifs des différentes sections du lycée et du collège. On peut alors refaire le même travail pour chaque section à partir des chiffres du ministère (même source) puis représenter côte à côte les sept graphiques en boîte.



On constate alors que les classes de collège sont assez semblables par leurs médianes (25) et par leurs dispersions, par contre l'effectif médian est très supérieur en lycée et spécialement en seconde, les classes de ce niveau présentent, de plus, les effectifs les plus élevés et ce de façon homogène alors que les premières et les terminales présentent plus de variations autour de la valeur centrale (effet des différentes séries du baccalauréat qui provoquent de petits effectifs dans les lycées de taille moyenne)[12].

Pour conclure, ce genre de travail peut donner du sens aux concepts de tendance centrale et de dispersion ainsi qu'à l'idée de comparaison de séries statistiques autrement que par le calcul des moyennes, ce qui n'a souvent pas de sens. En cela, c'est une bonne préparation au travail sur les paramètres de dispersion comme l'écart-type et l'écart interquartile qui seront vus en première.

III - Le hasard et la variabilité

L'étude de la statistique en collège devrait être une préparation à l'étude des probabilités au lycée. Si l'on veut que cette louable intention soit suivie d'effet, il faudrait que soient abordés au moins deux aspects qui constituent le cœur des problèmes où l'on fait intervenir des modèles probabilistes :

- la notion de hasard,
- la notion de variabilité des résultats de certaines expériences que l'on qualifie justement d'aléatoires, c'est-à-dire dont on ne peut prévoir, ni calculer le résultat.

La plupart des difficultés rencontrées plus tard en probabilités proviennent du passage de la réalité de l'expérience à la modélisation dans laquelle on effectuera les calculs. La première condition pour trouver un bon modèle est de bien définir l'épreuve aléatoire et par conséquent d'avoir une bonne représentation de ce qu'est une telle épreuve[13]. Il n'est pas évident de faire prendre conscience de la variabilité des résultats dans certaines expériences que l'on qualifie alors d'aléatoires. Pléonasme peut-être mais comment les élèves ne seraient-ils pas surpris que dans les mêmes conditions, une même expérience ne donne pas le même résultat. La physique (déterministe) a dû les convaincre que si les conditions initiales sont données, alors les résultats peuvent être calculés aux erreurs de mesure près !

---

[12] Le lecteur est invité à refaire cette étude avec des données plus récentes pour voir l'évolution dans les dernières années. Voir, par exemple, l'édition 2001 de l'ouvrage *Repères et références statistiques sur les enseignements et la formation*, Ministère de l'Education Nationale, Direction de l'Evaluation et de la Prospective.

[13] Bien que l'expression ne figure pas explicitement dans le programme officiel de seconde (et encore moins avant), une bonne représentation, à ce niveau, des épreuves aléatoires est un préalable à la construction d'une simulation correcte d'une expérience.

Il n’est pourtant pas difficile de trouver des contre-exemples sans revenir une fois de plus au lancer d’un dé !

- des graines de qualité semblable plantées en grande quantité dans un même champ produisent des plants de tailles différentes. On peut modéliser cette situation par une expérience aléatoire ;
- des frères et sœurs sont issus d'un même patrimoine génétique et pourtant il existe de nombreuses différences entre eux. Là encore, les lois de l’hérédité font intervenir le *hasard* comme *explication* ;
- de la même façon, on observera des variations entre des échantillons issus d’une même population car le hasard ne les aura pas constitués rigoureusement identiques. Par exemple, lorsque l’on étudie la variation de l’opinion par deux sondages successifs, on peut s’attendre à des résultats différents même s’il n’y a pas eu de modification au niveau de la population. C’est le rôle de la statistique inférentielle de faire la part de variation qui revient au hasard et celle qui traduit un réel changement de l’opinion. Le calcul des probabilités permet de calculer la probabilité d’un tel écart dans l’hypothèse où les deux échantillons seraient issus d’une même population, c’est-à-dire si l’opinion n’avait pas évolué. Si cette probabilité est trop petite (inférieure à 5 %, par exemple), le hasard, d'où découle la variabilité à laquelle on peut s’attendre dans la répétition d'une telle expérience, ne permet pas d’expliquer raisonnablement la différence observée et alors on refuse l’hypothèse d’une opinion stable. On parlera alors de différence significative (au niveau 5%).

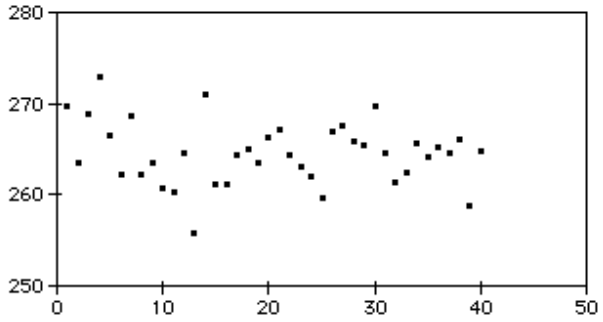
Il me semble que l’on peut faire en collège un travail d’approche de cette notion de variabilité, c’est-à-dire des variations des résultats dans une même épreuve aléatoire. Pour cela, on n’échappe à la manipulation de chiffres, ce qui peut paraître fastidieux mais qui me semble indispensable au moins une fois dans une scolarité si l’on veut mettre le doigt sur cette idée de variabilité. Il faudrait évidemment trouver des données, réelles si possible, et qui aient un sens pour les élèves. On peut en recueillir, par exemple, à l’occasion d’une visite dans une usine. Pour aller plus vite, on peut prendre un exemple dans un livre, mais les élèves vont-ils comprendre que sur une machine réglée de la même façon, avec la même matière première et à des instants très rapprochés (production en continu) les résultats obtenus puissent être différents et surtout que l’on ne puisse pas prévoir le suivant ?

Exemple : Les données suivantes[14] représentent le poids en grammes d'un joint d'étanchéité utilisé dans l'industrie automobile et obtenu d'une production continue.

269,7	263,4	268,8	272,9	266,4	262,2	268,7	262,3
263,6	260,7	260,3	264,5	255,8	271	261,2	261,2
264,4	265	263,4	266,2	267,1	264,4	263,1	262,1
259,7	267	267,6	265,9	265,5	269,8	264,6	261,4
262,4	265,6	264,1	265,3	264,5	266,1	258,7	264,8

Chaque valeur correspond à une production de 30 secondes. La variation dans l'écoulement du caoutchouc provenant de l'extrudeuse affecte directement les dimensions du joint. Quarante données ont été obtenues sur une période de production d'environ 30 minutes. Elles représentent un échantillon de la production.

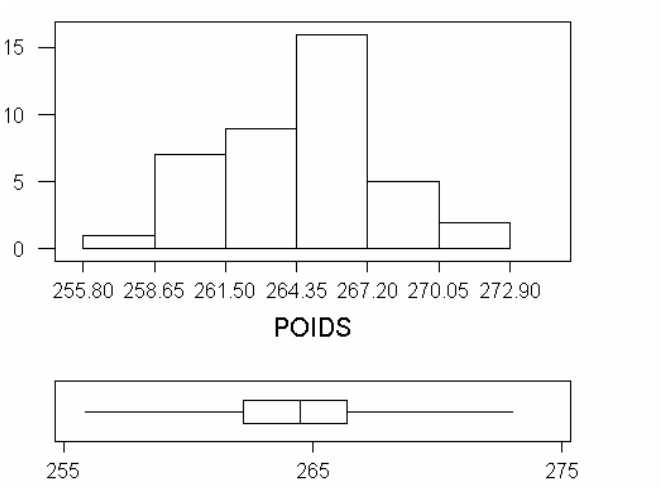
Les données sont dans l'ordre où elles ont été obtenues et peuvent être représentées dans cet ordre chronologique sur le graphique suivant :



Les valeurs semblent arriver au hasard, on ne peut prédire la suivante. Que peut-on faire ou dire dans ces conditions ? Pour dépasser les remarques naïves ou évidentes des élèves (« ça varie », « c'est pas très précis », « la machine n'est pas bonne ») et être efficace (mettre en place un contrôle de qualité, pouvoir dire quand il est nécessaire de régler la machine, savoir si un outil est adapté ou non à la production, etc.), on peut se placer dans un cadre statistique, c'est-à-dire dans un modèle expliquant, en partie par le hasard, la variabilité des résultats.

Une étude possible au collège pourrait commencer par une représentation graphique. On pensera bien sûr à l'histogramme puisque que les mesures individuelles sont variées (36 valeurs différentes sur 40).

[14] Les données de l'exemple sont extraites de BAILLARGEON, G. : *Maîtrise statistiques des procédés*, éditions SMG, Trois Rivières, Québec, 1992.



Explications : On a fait figurer en dessous de l'histogramme le graphique en boîte présenté au paragraphe précédent.

Mais les règles de construction de l'histogramme sont difficiles (le choix du nombre d'intervalles change la forme du graphique, problème des intervalles semi-ouverts, etc.) et, de plus, il peut ne pas avoir de sens pour les élèves.

On pourrait procéder, en guise de préalable, à la construction d'un graphique plus simple qui a l'avantage de donner à peu près la même représentation de la série et cela en ne perdant aucune information. Il s'agit du graphique en tige et feuilles[15] (ou stem and leaf).

Présentation d'un graphique tige et feuilles :

La valeur 255,8 est représentée par

255	8
tige	feuille

De ce graphique, on peut faire ressortir quelques observations.

La distribution des valeurs n'est pas quelconque, encore moins uniforme. On a beaucoup de *chances* de se trouver proche de la valeur médiane qui est 264,5. On remarque que 31 valeurs (soit plus de 75 %) se trouvent entre 261,2 et 267,6. On retrouve un peu d'ordre dans le hasard.

255 8
256
257
258 7
259 7
260 37
261 224
262 1234
263 1446
264 1445568
265 03569
266 124
267 016
268 78
269 78
270
271 0
272 9

[15] John W. TUKEY, op. cit. Pour plus de détails sur ce type de représentation, voir l'article d'Hubert RAYMONDAUD : *Quelques pièges de la description d'une série statistique* dans ce même volume.

Prolongements possibles (compréhensibles au collège)

- Que peut-on conclure de l'observation de ces 40 valeurs si on suppose la machine réglée convenablement au début de la production ?
- Que penser, si dans la suite de la production, une heure plus tard par exemple, on mesure une valeur de 268,5 ?
- Que penser, si dans la suite de la production, on mesure une valeur de 255 ?
- A partir de quelle(s) valeur(s) doit-on suspecter un déréglage de la machine ?
- Que penser de la machine utilisée si on doit avoir impérativement un poids compris entre 264 et 266 pour que la production soit acceptable ?

IV - Remarques et conclusion

Il s'agit seulement de faire prendre conscience de quelques faits :

- dans de nombreuses expériences, répétées pourtant dans les mêmes conditions, les résultats présentent une certaine variabilité,
- les mathématiques prennent en compte ce genre de situations en fournissant des modèles faisant intervenir le hasard. ; on peut alors retrouver une certaine stabilité au cœur de ces variations et faire des prévisions,
- une valeur éloignée de la moyenne ou de la médiane n'est pas impossible mais doit attirer notre attention,
- dans le cas étudié, la variabilité des résultats est liée à la précision de la machine c'est-à-dire sa capacité à produire des pièces dont la mesure est plus ou moins proche de la valeur de réglage.

Ce genre de travail devrait permettre aux élèves :

- d'être confronté à une épreuve aléatoire concrète,
- d'appréhender les effets du hasard,
- de retrouver des régularités au sein des résultats aléatoires,
- d'analyser une série statistique de façon critique,
- de sensibiliser les élèves à une application de la statistique, le contrôle de qualité[16,]
- et surtout de montrer que l'on ne fait pas de la statistique uniquement pour obtenir un beau graphique ou une moyenne avec quatre décimales.

L'objectif de cet article était d'illustrer ce que l'étude de la statistique peut apporter à la formation générale ainsi qu'aux autres domaines des mathématiques, à

[16] Dans la réalité, les contrôles sont effectués à partir de la moyenne des valeurs d'un échantillon dont l'effectif est souvent égal à 5. Voir, par exemple, Gérald BAILLARGEON, op. cit.

la préparation à l'étude des probabilités et à la formation du citoyen. On a également insisté sur la familiarisation avec l'idée de variabilité qui relativise l'importance quasi mystique donnée à la moyenne et montre la nécessité de prendre en compte la dispersion d'une série statistique[17].

S'il présente beaucoup d'intérêt, cet enseignement présente aussi de nombreuses difficultés. La réflexion doit être poursuivie dans différentes directions, par exemple sur la pertinence de l'introduction des probabilités (au moins des expériences aléatoires) au collège[18], sur l'apprentissage des pourcentages et sur celui de la lecture de graphiques (Quelles sont les conceptions spontanées des élèves devant un graphique ? Comment les aider à construire de bonnes images mentales ?). Cela ne se fera qu'en réfléchissant à des exemples concrets et intéressants d'utilisation qui donnent, en prime, du sens aux concepts statistiques étudiés à ce niveau de scolarité[19].

[17] Cette idée est développée dans l'article de GIRARD, J. C. : A bas la moyenne !, *Repères-IREM n° 33*, octobre 1998.

[18] Pour des exemples, voir l'article de GIRARD, J. C., HENRY, M., PICHARD, J. F., PARZYSZ, B. : Quelle place pour l'aléatoire au collège ?, *Repères-IREM* n° 42, janvier 2001.

[19] Voir par exemple, ROBERT C. : *L'empereur et la girafe - Leçons élémentaires de statistiques*, Diderot Editeur, Paris, 1995.

Quartiles, déciles et tutti quantiles

Jean Claude GIRARD

La définition des quartiles et des déciles proposée dans les nouveaux programmes de première peut paraître surprenante au premier abord :

« *Le premier (respectivement le troisième) quartile est le plus petit élément q_1 (respectivement q_3) des valeurs des termes de la série ordonnée par ordre croissant, tel qu'au moins 25 % (respectivement 75 %) de ces valeurs soient inférieures ou égales à q_1 (respectivement q_3).*

Le premier décile (respectivement le deuxième, le troisième, etc.) est le plus petit élément d_1 (respectivement d_2, d_3, etc.) des valeurs des termes de la série ordonnée par ordre croissant, tel qu'au moins 10 % (respectivement 20 %, 30 %, etc.) des valeurs soient inférieures ou égales à d_1 (respectivement d_2, d_3, etc.). »[1]

Elle n'est pas sans poser de problèmes didactiques. Or, il existe d'autres définitions possibles et les calculatrices, les tableurs et les logiciels de statistique sont très divergents sur le sujet.

Après avoir rappelé le cas des variables aléatoires continues pour lesquelles les notions de quartiles, déciles et plus généralement quantiles sont définies de manière non ambiguë, plusieurs définitions seront proposées pour les séries statistiques (nécessairement discrètes).

Enfin, d'une façon plus générale, l'utilité des quantiles sera illustrée d'abord dans la comparaison de deux distributions statistiques, puis dans la comparaison d'une série observée à une distribution théorique (test de normalité par exemple).

I - Quantiles d'une loi de probabilité continue à densité non nulle sur un intervalle

Si une variable aléatoire réelle X suit une loi continue de densité de probabilité f non nulle sur un intervalle I, la probabilité que X prenne une valeur comprise entre a et b est donnée par : $P(a < X < b) = \int_a^b f(t)\,dt$.

F, la fonction de répartition de X est définie par : $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$.

Dans notre hypothèse, F est continue et strictement croissante sur I, donc

[1] GEPS : Document d'accompagnement des programmes de premières.

inversible. Le quantile d'ordre α de cette loi est le réel q_α appartenant à I tel que $F(q_\alpha) = \alpha$. Autrement dit : $q_\alpha = F^{-1}(\alpha)$.

Par exemple :

- pour la médiane[2] : $Me = q_{0,5} = F^{-1}(0,5)$;
- pour les quartiles : $Q_1 = q_{0,25} = F^{-1}(0,25)$, $Q_3 = q_{0,75} = F^{-1}(0,75)$ et on a $Q_2 = Me$;
- pour les déciles : $D_1 = q_{0,1} = F^{-1}(0,1)$, $D_2 = q_{0,2} = F^{-1}(0,2)$, etc. et on a $D_5 = Me$.

II - Premier exemple : la loi exponentielle

La loi exponentielle modélise les durées de vie sans vieillissement.

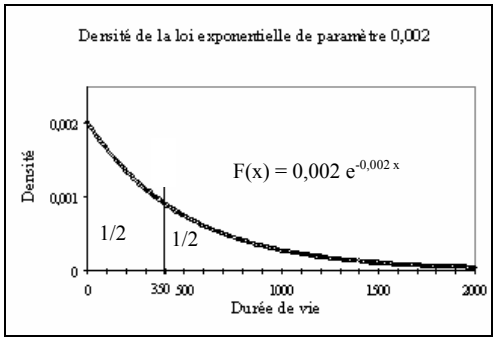
La loi exponentielle de paramètre $\lambda > 0$ a pour densité de probabilité la fonction f définie sur $\mathbb{R}^+$ par $f(x) = \lambda.e^{-\lambda x}$.

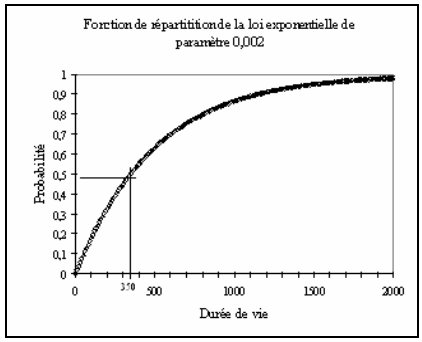
Si une variable aléatoire X suit la loi exponentielle de paramètre λ, sa fonction de répartition est la fonction F, nulle sur $\mathbb{R}^-$ et définie sur $\mathbb{R}^+$ par $F(x) = P(X \leq x) = 1 - e^{-\lambda x}$. L'espérance mathématique de cette loi est $E(X) = \frac{1}{\lambda}$.

La médiane (demi-vie) vérifie $1 - e^{-\lambda Me} = \frac{1}{2}$, d'où $Me = \frac{1}{\lambda} \ln 2$.

Exemple numérique :

Si la loi de la durée de vie (en heures) X d'un certain composant est modélisée par la loi exponentielle de paramètre $\lambda = 0,002$ on a $E(X) = \frac{1}{0,002} = 500$ h et $Me \approx 350$ h.





[2] A ne pas confondre avec la définition de la *médiane* empirique d'une série statistique donnée en seconde :
« *On ordonne la série des observations par ordre croissant ; si la série est de taille 2 n + 1, la médiane est la valeur du terme de rang n + 1 dans cette série ordonnée ; si la série est de taille 2n, la médiane est la moyenne des valeurs des termes de rang n et n + 1 dans cette série ordonnée.* »
(Document d'accompagnement des programmes de première).

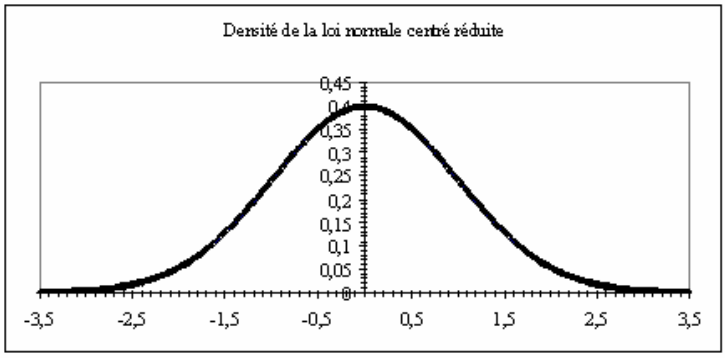
III - Deuxième exemple : la loi normale ou de Gauss[3]

La loi normale d'espérance mathématique μ et d'écart-type σ, notée $N(\mu, \sigma)$, a pour densité de probabilité la fonction f définie pour tout x réel par :

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

.

Si une variable aléatoire X suit cette loi, la variable centrée réduite $\frac{X-\mu}{\sigma}$ suit la loi normale centrée réduite, (d'espérance mathématique 0 et d'écart-type 1), de densité ϕ donnée par $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$.

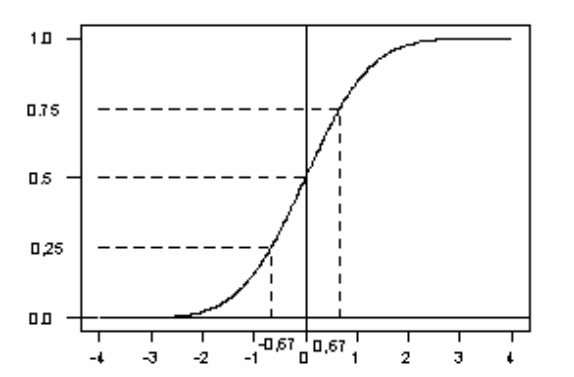


La loi normale modélise de nombreux phénomènes (pas tous !).

L'intérêt de savoir si une distribution observée peut être considérée comme un échantillon gaussien (c'est-à-dire en fait si on peut la considérer comme une série de valeurs résultant d'expériences indépendantes de même loi normale) est que l'on peut alors faire des calculs (de probabilité) à partir de cette loi « théorique ».

Par symétrie, la médiane Me est nulle. Par lecture des tables (ou avec une calculatrice), on trouve les autres quartiles $Q_1 \approx -0{,}67$ et $Q_3 \approx 0{,}67$ représentés ci-dessous :

[3] Sur la dénomination de cette loi, voir l'article de J.-F. PICHARD : *Théorie des erreurs, courbe en cloche et normalité* dans ce même volume.



Fonction de répartition de la loi normale centrée réduite :

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

On obtient pour les déciles :

i	1	2	3	4	5	6	7	8	9
d_i	– 1,28	– 0,84	– 0,52	– 0,25	0	0,25	0,52	0,84	1,28

IV - Quantiles pour une série statistique

Il est souvent utile, si le contexte expérimental le permet[4], de considérer les données d'une série statistique observée comme des valeurs résultant d'expériences aléatoires indépendantes. Ces valeurs peuvent être représentées par les réalisations de variables aléatoires dont la loi commune (théorique) est continue.

Mais les données d'une série statistique sont naturellement toujours en nombre fini. Présenter les données sous forme classées (par intervalles) introduit une perte d'information que l'on compense par des conventions[5] sur la distribution des valeurs possibles de la variable théorique modèle dans chaque intervalle. Cependant, la détermination des quantiles, en particulier des quartiles et des déciles, doit être faite sur des séries de données statistiques observées (non regroupées en classes).

A partir d'une série statistique $(x) = (x_i)_{1 \le i \le n}$, on définit la fonction de répartition empirique $\hat{F}_{(x)}$ qui à tout t réel donné, fait correspondre la fréquence des valeurs observées de la série inférieures ou égales à t : $\hat{F}_{(x)}(t) = \frac{\text{Card}\ \{x_i / x_i \le t\}}{n}$.

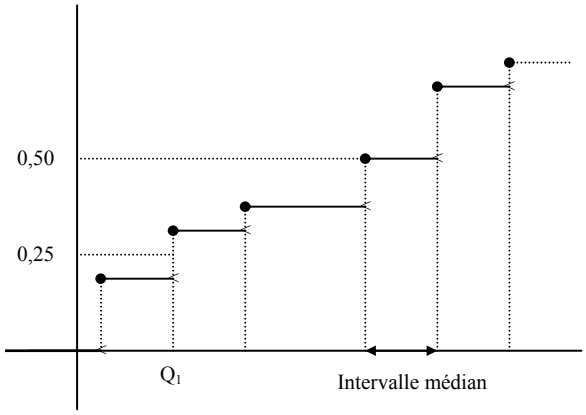
$\hat{F}_{(x)}$ est une fonction en escalier qui n'a pas de fonction réciproque[6] et on ne peut pas définir, par exemple, le premier quartile par $Q_1 = \hat{F}_{(x)}^{-1}(0{,}25)$. Si la valeur α (0,25 ci-

[4] Cf. *Théorie des erreurs, courbe en cloche et normalité*, article cité.

[5] On considère une distribution uniforme des valeurs dans chaque classe pour déterminer les quantiles de la série de données, mais on suppose que toutes les valeurs sont regroupées aux *centres des classes* pour le calcul de la moyenne et de l'écart-type empiriques.

[6] Conventionnellement, $\hat{F}_{(x)}^{-1}(\alpha)$ désigne l'ensemble des réels ayant pour image α par $\hat{F}$.

dessous) n'est pas une valeur prise par $\hat{F}_{(x)}$, $\hat{F}_{(x)}^{-1}(\alpha)$ est vide ; sinon (cas de 0,50), l'image réciproque de α par $\hat{F}_{(x)}$ est un intervalle (intervalle médian).



Fonction de répartition empirique $\hat{F}_{(x)}$

La définition des quartiles retenue pour les programmes des classes de première (citée en début d'article) conduit à prendre pour le premier quartile, la plus petite valeur Q_1 de la série telle que $\hat{F}_{(x)}(Q_1) \geq 0{,}25$ comme indiqué sur le graphique ci-dessus.

On peut démontrer que si la série statistique est constituée par des valeurs prises par un échantillons de n variables aléatoires X_i indépendantes et de même loi, les quantiles empiriques sont de *bons* estimateurs des quantiles de cette loi[7]. De surcroît, le théorème de Glivenko-Cantelli (théorème fondamental de la statistique) établit que pour n assez grand, la fréquence cumulée empirique $\hat{F}_n$ des valeurs obtenues dans un échantillon aléatoire de taille n d'une loi théorique est un *bon* estimateur de la fonction de répartition F de cette loi[8].

---

[7] Par « bons estimateurs », on sous-entend que les quantiles empiriques ainsi déterminés convergent en un certain sens (i.e. presque sûrement) vers les quantiles correspondants de la loi des X_i. (On dit aussi que ces estimateurs sont « consistants »).

[8] PluS précisément : Pour presque toute suite infinie $(x) = (x_i, \ldots, x_n, \ldots)$ d'observations indépendantes (i. e. sauf pour un ensemble de suites de probabilité nulle), notant $(x|_n)$ la suite finie des n premiers termes de (x), on a : $\lim_{n \to +\infty} \sup_{t \in \mathbb{R}} \left| \hat{F}_{(x|_n)}(t) - F(t) \right| = 0$ (convergence presque uniforme de la suite des fonctions de répartition empiriques vers la fonction de répartition de la loi théorique).

V - Remarques d'ordres pratique et didactique

La définition donnée ci-dessus est donc satisfaisante du point de vue mathématique. Ce n'est par contre peut-être pas la meilleure du point de vue didactique. En effet :

- On ne peut passer outre la difficulté de lecture de la définition : combiner *plus petit*, *au moins*, *inférieur ou égal* dans une même phrase (surtout pour une définition) n'est pas la meilleure façon d'être clair.
- Cette définition conduit d'autre part à ce que Q_2, c'est-à-dire le deuxième quartile ne soit pas égal à la médiane[9]. Ce n'est pas en anticipant « *qu'il n'y a pas de raison de signaler qu'avec la définition adoptée, la médiane n'est pas le second quartile, sauf si un élève pose précisément la question.* », comme cela figurait dans une version provisoire du document d'accompagnement des programmes de premières, que l'on élimine cette curiosité.
- Elle ne vérifie pas les propriétés de symétrie. Par exemple, le rang de Q_1 par ordre croissant est différent du rang de Q_3 par ordre décroissant, c'est-à-dire : si Q_1 est la $k^{ième}$ valeur à partir du minimum, Q_3 n'est pas nécessairement la $k^{ième}$ valeur à partir du maximum. De plus, $\frac{\text{rang}(Q_1) + \text{rang}(Q_3)}{2} \neq \text{rang(Me)}$ et même $\frac{\text{rang}(Q_1) + \text{rang}(Q_3)}{2} \neq \text{rang}(Q_2)$
- Ceci ne concourt pas à donner une idée claire du concept de paramètre de position. En particulier l'idée que les quartiles partagent la série en 4 sous-séries de mêmes effectifs n'apparaît pas immédiatement.

En conclusion : La définition adoptée pour les quartiles et les déciles n'est peut-être pas la meilleure pour une découverte de ces notions puisqu'il existe d'autres définitions effectivement utilisées par les statisticiens. Le paragraphe suivant présente trois autres façons courantes de définir les quartiles d'une série statistique (parmi une multitude de possibilités).

VI - Autres façons de définir les quartiles

L'idée la plus simple quand on a partagé une série ordonnée en deux sous-séries d'égal effectif par une valeur médiane, c'est de repartager chacune des sous-séries à nouveau en deux parties d'égal effectif. Différents cas se présentent pour obtenir ces deux sous-séries selon la parité de la taille n de la série donnée : si n est impair, on exclut la donnée de rang $\frac{n+1}{2}$, si n est pair on conserve toutes les données. On a alors deux options : soit on ne reprend pas la médiane dans chacune des deux sous-séries (c'est la méthode employée par les calculatrices Casio ou Texas Instruments,

[9] La définition de la médiane, donnée en seconde, est rappelée en note 2.

par exemple) soit on la reprend (c'est la méthode employée par Excel). Une troisième méthode généralisable à tous les quantiles est employée par certains logiciels de statistique (c'est la méthode employée par MINITAB).

1 - Méthode des calculatrices

On suppose que les n données de la série sont ordonnées dans l'ordre croissant : x_1, x_2,…, x_n. Si n est impair, la médiane est égale à la donnée de rang $\frac{n+1}{2}$ occupant la place centrale. Si n est pair, la médiane est égale à la moyenne entre les deux données centrales de rangs $\frac{n}{2}$ et $\frac{n}{2}+1$. Par abus, on dit encore que c'est la valeur (fictive) de « rang » $\frac{n+1}{2}$.

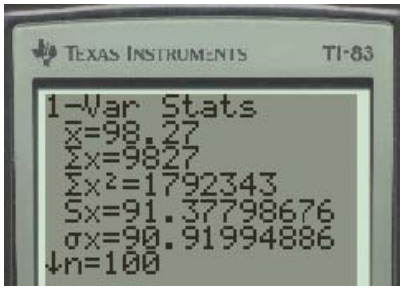
Pour définir les quartiles, on procède de même pour les deux sous-séries obtenues en excluant la donnée centrale quand n est impair, sans rien exclure si n est pair.

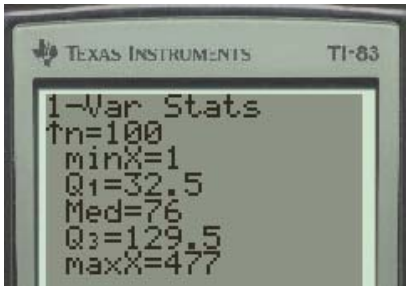
Exemple : Les données suivantes représentent des durées de vie (en heures) d'un certain matériel.

1	4	4	5	5	7	7	10	10	11
12	13	14	15	19	20	20	21	21	23
23	26	28	32	32	33	33	34	34	38
39	41	43	46	46	47	49	50	52	52
59	59	66	66	71	71	73	74	75	75
77	77	82	82	84	86	87	92	92	97
102	103	103	107	108	108	110	111	111	119
121	121	123	125	129	130	136	138	139	141
143	145	148	163	167	181	185	192	206	214
233	254	259	266	275	294	296	374	405	477

$n = 100$, donc $\frac{n+1}{2} = 50{,}5$ et $Me = \frac{x_{50} + x_{51}}{2} = \frac{75+77}{2} = 76.$

$Q_1 = \frac{x_{25} + x_{26}}{2} = \frac{32+33}{2} = 32{,}5 \quad Q_3 = \frac{x_{75} + x_{76}}{2} = \frac{129+130}{2} = 129{,}5.$





2 - Méthode d'Excel

On inclut cette fois-ci la médiane, valeur de « rang $\frac{n+1}{2}$ », dans les deux sous-séries pour le calcul des quartiles. Le minimum de la première sous-série est la donnée de rang 1, le maximum est la valeur de « rang $\frac{n+1}{2}$», donc la médiane de la première sous-série est la valeur de « rang » $\frac{1}{2}\left(1+\frac{n+1}{2}\right)$, soit $\frac{n+3}{4}$. Si ce « rang » n'est pas entier, il ne correspond pas à une valeur de la série et on fait une interpolation linéaire sur les valeurs adjacentes pour obtenir cette médiane, premier quartile Q_1 de la série donnée.

Application à l'exemple des durées de vie

$n = 100$, $\frac{n+3}{4} = 25{,}75$ donc $Q_1 = x_{25} + 0{,}75\,(x_{26} - x_{25}) = 32 + 0{,}75\,(33 - 32) = 32{,}75$.

Le minimum de la deuxième sous-série est la valeur de « rang » $\frac{n+1}{2}$, le maximum est la valeur de rang n, donc la médiane de la deuxième sous-série est la valeur de « rang » $\frac{1}{2}\left(\frac{n+1}{2}+n\right)$, soit $\frac{3n+1}{4}$. Si ce « rang » ne correspond pas à une valeur de la série, on fait une interpolation linéaire pour obtenir le quartile Q_3.

Application à l'exemple des durées de vie

$n = 100$, $\frac{3n+1}{4} = 75{,}25$ $Q_3 = x_{75} + 0{,}25\,(x_{76} - x_{75}) = 129 + 0{,}25\,(130 - 129) = 129{,}25$.

C'est la méthode employée par J. W. TUKEY pour construire les premières boîtes à moustaches (box and whiskers plot[10]).

Durée de vie

	R	S	T	U
1	1			
2	4			
3	4		minimum	1
4	5		Q1	32,75
5	5		Me	76
6	7		Q3	129,25
7	7		Maximum	477
8	10			
9	10			
10	11		effectif	100
11	12			
12	13		moyenne	98,27
13	14			
14	15		écart type	91,3779868
15	19		écart type p	90,9199489

[10] *Exploratory Data Analysis*, Addison-Wesley, 1977

3 - Méthode de MINITAB (et d'autres[11])

Les quartiles sont les valeurs de « rangs » $\frac{k\,(n+1)}{4}$, pour $1 \le k \le 3$. Q_1 est donc la valeur de « rang » $\frac{n+1}{4}$, Q_2 la valeur de « rang » $\frac{n+1}{2}$, c'est-à-dire la médiane, et Q_3 la valeur de « rang » $\frac{3\,(n+1)}{4}$. Un « rang » non entier ne correspond pas à une valeur de la série, le quartile correspondant est alors obtenu par une interpolation linéaire entre les deux valeurs de rangs entiers voisins.

Application à l'exemple des durées de vie

$n = 100$, $\frac{n+1}{4} = 25{,}25$ donc $Q_1 = x_{25} + 0{,}25\,(x_{26} - x_{25}) = 32 + 0{,}25\,(33 - 32) = 32{,}25$.

$\frac{n+1}{2} = 50{,}5$ donc $Me = Q_2 = x_{50} + 0{,}5\,(x_{51} - x_{50}) = 76$.

$\frac{3\,(n+1)}{4} = 75{,}75$ donc $Q_3 = x_{75} + 0{,}75\,(x_{76} - x_{75}) = 129 + 0{,}75\,(130 - 129) = 129{,}75$.

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev
C10	100	98.27	76.00	88.47	91.38

Variable	Min	Max	Q1	Q3
C10	1.00	477.00	32.25	129.75

La même définition est employée pour un quantile quelconque, par exemple les déciles sont les valeurs des observations de « rangs » $\frac{k\,(n+1)}{10}$ pour $1 \le k \le 9$.

Par exemple : pour $n = 100$, $D_1 = x_{10} + 0{,}1(x_{11} - x_{10})$.

Remarque 1 : Dans les trois cas, les rangs de Q_1 et Q_3 sont symétriques par rapport au rang de la médiane, on a $\frac{\text{rang}(Q_1) + \text{rang}(Q_3)}{2} = \text{rang}(Q_2) = \text{rang}(Me)$, donc $Me = Q_2 = D_5$. Cependant les valeurs obtenues pour les quartiles peuvent être différentes. Ainsi, dans l'exemple des durées de vie, pour Q_3, on obtient 129,5 avec les calculatrices, 129,25 avec Excel et 129,75 avec Minitab.

Remarque 2 : Les commentaires des programmes de premières précisent que si le nombre d'observations est grand, les valeurs des quartiles sont peu différentes quelle que soit la méthode utilisée pour les calculer. Ceci n'est pas tout à fait vrai, les valeurs peuvent être notablement différentes. Ce qui compte, ce n'est pas tant le nombre d'observations que la densité de ces observations dans la zone où l'on

[11] Voir par exemple : *Dictionnaire encyclopédique de Statistique*, Yadolah DODGE, Dunod, 1993.

calcule les quartiles. En effet, les quartiles, et en particulier la médiane, ne sont pas sensibles aux valeurs extrêmes (contrairement à la moyenne), mais ils sont sensibles, par contre, aux discontinuités. Ceci signifie que le calcul des quartiles peut ne pas être pertinent dans certains cas (comme le calcul de la moyenne n'est pas pertinent quand la série comporte des valeurs très éloignées les unes des autres).

Remarque 3 : Tous ces calculs n'ont d'intérêt que si l'on utilise les résultats obtenus ! Le paragraphe suivant se propose de montrer une utilisation possible.

VII - Comparaison de deux séries empiriques

Dans la pratique quotidienne, la comparaison de deux séries se fait souvent à partir de leurs seules moyennes. Des moyennes égales traduisent pourtant des situations diverses. Même l'utilisation d'une mesure de dispersion comme l'écart-type ne suffit pas, car on peut construire des distributions de formes très différentes ayant la même moyenne et le même écart-type[12]. L'utilisation des quantiles (des déciles, par exemple) permet de parvenir à des résultats plus fins dans la comparaison de deux distributions.

Le graphique suivant (en tige et feuilles) donne les répartitions (par ordre croissant et non par ordre chronologique) du nombre de jours de neige à Paris dans les deux moitiés du 20ème siècle. Par exemple, le maximum dans la première moitié du siècle a été de 34 jours de neige et de 36 dans la deuxième moitié, deux années ont eu 17 jours de neige entre 1900 et 1948, alors qu'il n'y en a eu qu'une seule entre 1949 et 1997.

Années 1900-1948		Années 1949-1997
1	0	122244
98877776665	0	55555777899
4444433322211100	1	01222222233344
998887765	1	6678889
30000	2	01222234
996	2	9
4221	3	1
	3	6

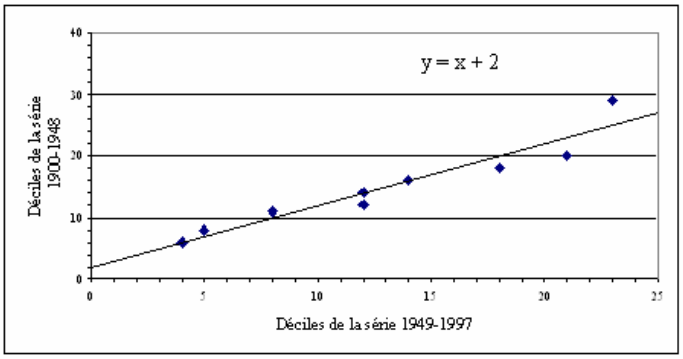
Les deux séries ont 49 valeurs. Les déciles[13] sont les valeurs de rangs 5, 10, 15, 20, 25, 30, 35, 40 et 45, c'est-à-dire :

[12] C'est un exercice très formateur ! Cf. l'article d'Hubert RAYMONDAUD : *Quelques pièges de la description d'une série statistique*.

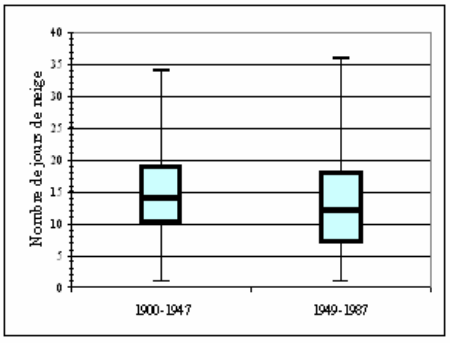
[13] Avec la définition du programme des classes de première.

Déciles	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9
Série 1900-1948	6	8	11	12	14	16	18	20	29
Série 1949-1997	4	5	8	12	12	14	18	21	23

Si les deux séries étaient identiques, elles auraient les mêmes déciles. Si on reporte dans un repère orthonormé les déciles de la deuxième série en abscisse et ceux de la première en ordonnées, on devrait obtenir des points alignés sur la première bissectrice[14]. Ce n'est pas le cas, les points sont plutôt alignés sur la droite d'équation $y = x + 2$.



Ceci signifie qu'il a eu globalement deux jours de neige en moins chaque année dans la deuxième moitié du siècle. Le graphique en boîte des deux séries confirme ce résultat.



[14] Plus généralement, si Q est un quantile de la série des valeurs de X, a Q + b est le quantile correspondant de la série a X + b.

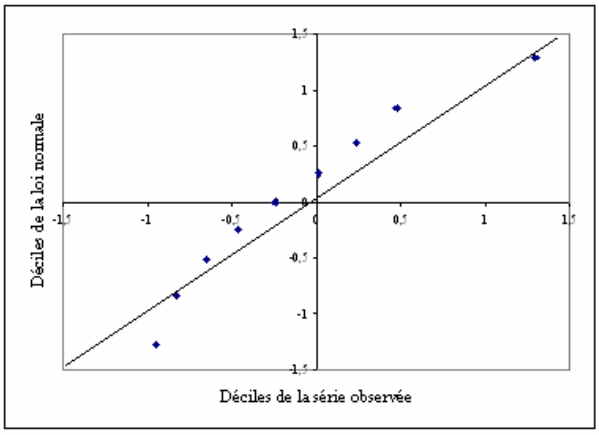
VIII - Contrôle de normalité

Etant donné une série statistique observée, on peut être amené à s'interroger sur la validité du modèle suivant lequel les données seraient des valeurs prises par des variables aléatoires indépendantes et de même loi théorique connue. Par exemple, certaines techniques statistiques ne sont valables que dans le cas de variables normales, on doit donc s'assurer que ce modèle est acceptable. Pour cela il faut (au moins) que les déciles de la loi théorique et ceux de la série observée soient à peu près égaux. La même représentation graphique que précédemment doit fournir des points alignés sur la première bissectrice. Dans le cas contraire, on ne retiendra pas ce modèle.

Exemple : En reprenant les données de l'exemple sur les durées de vie, peut-on admettre l'hypothèse de normalité ?

On calcule les déciles de la série observée (centrée réduite) d que l'on compare avec les déciles théoriques d' de la loi normale (voir II).

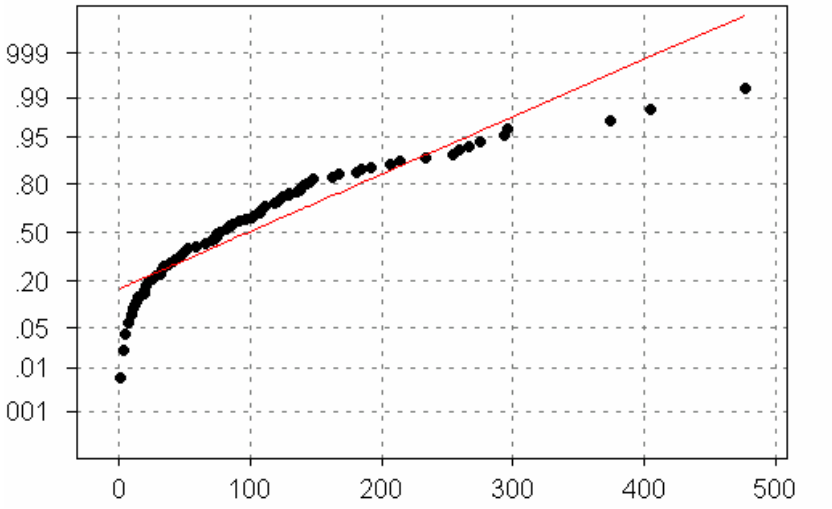
d_i	– 0,95	– 0,83	– 0,66	– 0,46	– 0,24	0,01	0,23	0,47	1,29
d'_i	– 1,28	– 0,84	– 0,52	– 0,25	0	0,25	0,52	0,84	1,28



Il semble que l'hypothèse de normalité soit à rejeter.

Remarque :

On obtient le même résultat en reportant toutes les données sur du papier gausso-arithmétique[15].



Bibliographie

SAPORTA, G., *Probabilités, Analyse des données et Statistique*, Technip, 1990.

CHAMBERS, J. M., CLEVELAND W. S., KLEINER B. TUKEY P. A., *Graphical Methods for Data Analysis*, Chapman & Hall, 1983.

[15] Voir par exemple : ABBOUD, N., AUDROING, J. F. : *Probabilités et inférence statistique*, Nathan, 1989.

Photo publiée avec l'aimable autorisation de la brasserie de Saint-Sylvestre (59)

Cuve de fermentation

Quelques pièges de la description d'une série statistique

Hubert RAYMONDAUD

I - Introduction

Cet article a pour objectif de montrer les insuffisances de la moyenne et de l'écart-type pour décrire une série statistique à une variable quantitative. Il présente l'étude d'un exemple concret et peut être utilisé à partir de la classe de seconde. Il propose, bien sûr, en même temps, une démarche et des outils : *tige et feuilles*, diagrammes, résumé en chiffres, boîte à pattes, permettant une description suffisante pour une première analyse des données qu'on peut appeler analyse exploratoire.

La description d'une série statistique à une variable quantitative est faite pour synthétiser et visualiser l'information contenue dans les données. Qui dit synthèse et visualisation, dit donc graphisme, judicieusement choisi pour rendre compte de cette information. C'est ce qui a suscité le travail remarquable de TUKEY (1977)[1]. Il nous propose ses outils semi-graphiques et graphiques dont nous allons mettre certains en pratique.

Mettre des outils en pratique, d'accord, mais sur quel matériau et pour quoi faire ? C'est un peu comme une lunette astronomique : avant de l'utiliser, il faut savoir qu'elle est faite pour voir les objets célestes de plus près. Si l'on veut regarder les nébuleuses d'Orion, il faut la diriger vers le bon endroit dans la voûte céleste, sous peine de voir, certes, plein de choses très intéressantes, mais pas l'objet que l'on s'était fixé au départ. La statistique descriptive est composée d'outils, qu'il faut savoir utiliser à bon escient, autrement dit, avec une méthode.

Nous proposons de traiter un cas concret, qui sert de support pour introduire une méthode plus générale[2].

---

[1] Pour plus de détails sur les techniques de l'analyse exploratoire, on pourra consulter les références suivantes : TUKEY J. W. (1977), HAOGLIN et al. (1983), CHAMBERS J. M. et al (1983) et J.-C. GIRARD (1993) ainsi que son article *Quartiles, déciles et tutti quantiles* dans ce même volume.

[2] Un exemple détaillé de description et de comparaison de séries univariées, à l'aide d'un tableur, figure dans l'article *Réinvestir le vocabulaire et quelques méthodes de la description des séries univariées quantitatives* (*Panorama A*), à paraître dans le volume 2 de *Statistique au lycée.*

II - Description d'une série statistique à une variable quantitative : un exemple, une méthode

Afin d'évaluer l'efficacité du nettoyage de trois de ses cuves de fermentation, une brasserie fait un contrôle en mesurant la qualité bactériologique des eaux de leur dernier rinçage. Pour chaque cuve contrôlée, on effectue 50 prélèvements à partir du tuyau de vidange, à 50 moments différents d'une même vidange. Les résultats de ces prélèvements dans les trois cuves, sont donnés dans les tableaux qui suivent (document 1) en nombre moyen de *bactéries totales* par ml (bact./ml), classés par ordre croissant[3]. *Bactéries totales* signifie qu'on dénombre toutes les espèces de bactéries, par opposition à d'autres méthodes où l'on se limite à certaines espèces seulement.

CUVE 1

1	34,0578	11	98,1269	21	136,7880	31	171,8743	41	231,0180
2	39,8810	12	101,0069	22	138,3488	32	189,3360	42	234,3610
3	43,3432	13	106,2150	23	140,7957	33	191,0610	43	236,2030
4	45,8474	14	109,2210	24	142,4932	34	191,3664	44	237,0360
5	61,7631	15	117,7186	25	143,3407	35	195,5220	45	258,2447
6	65,6779	16	118,1281	26	147,2411	36	196,7330	46	270,4860
7	75,7151	17	119,4790	27	149,0102	37	200,7820	47	282,1780
8	89,6167	18	125,4248	28	149,7260	38	203,4700	48	286,2670
9	89,8100	19	126,0350	29	156,1123	39	205,2845	49	292,4440
10	96,3790	20	135,4515	30	156,8513	40	221,5461	50	307,2880

CUVE 2 :

1	72,2714	11	102,2699	21	128,8510	31	153,5820	41	191,4610
2	76,5449	12	106,9700	22	132,5062	32	154,9971	42	210,4620
3	78,9124	13	107,1120	23	135,1847	33	155,7972	43	219,0080
4	80,2609	14	110,3013	24	136,7666	34	167,3450	44	229,2370
5	83,4694	15	115,2728	25	136,8911	35	171,7649	45	261,5920
6	88,1316	16	118,6713	26	140,6101	36	177,9620	46	263,2460
7	90,3279	17	118,7085	27	143,6863	37	179,5180	47	289,2660
8	93,3724	18	123,8917	28	147,5271	38	180,0030	48	325,4030
9	94,7748	19	126,2213	29	151,0397	39	182,6480	49	352,6700
10	100,0937	20	128,3722	30	153,4499	40	185,2490	50	388,4300

[3] On pourra noter que, dans cet exemple, les quantités observées ne peuvent être considérées comme des issues d'expériences aléatoires indépendantes. La modélisation probabiliste n'est alors pas élémentaire.

CUVE 3 :

1	44,4536	11	77,8625	21	126,6420	31	201,9560	41	227,8030
2	55,3463	12	78,1565	22	157,5640	32	206,7090	42	228,8810
3	59,3543	13	78,8891	23	160,4500	33	206,7310	43	232,1380
4	66,5909	14	82,6468	24	167,6850	34	209,6590	44	234,0580
5	69,6085	15	85,5969	25	174,9600	35	210,8570	45	234,1570
6	72,1987	16	85,7852	26	182,0850	36	211,9280	46	240,8630
7	73,0620	17	93,2374	27	189,7710	37	214,8260	47	245,7610
8	74,6154	18	94,7565	28	192,1420	38	215,1860	48	248,2070
9	76,7215	19	99,4083	29	197,8490	39	222,3270	49	253,8790
10	77,7942	20	101,9530	30	201,0010	40	225,1600	50	292,8670

Document 1 : Qualité du nettoyage mesurée par le nombre de bactéries totales par ml dans l'eau du dernier rinçage

Si pour décrire les trois séries statistiques présentées, correspondant aux mesures effectuées à partir des trois cuves, nous nous bornons à calculer leurs moyennes et leurs écarts-types, comme on en a malheureusement trop souvent la mauvaise habitude, nous obtenons les mêmes valeurs : 157,24 bact./ml pour les moyennes et 70,52 bact./ml pour les écarts-types.

Ne nous hâtons surtout pas de conclure, mais efforçons-nous de faire une description digne de ce nom.

L'objectif de cette description est de déterminer, d'illustrer et de comparer les distributions observées pour chacune des cuves du caractère prenant pour valeur à chaque prélèvement le nombre de bactéries totales par ml. La forme des distributions observées peut nous aider à formuler certaines hypothèses comparatives quant aux populations microbiennes ayant habité ces trois cuves, ainsi que sur la qualité des rinçages.

Pour décrire nos trois séries, nous proposons trois outils graphiques :

- tiges et feuilles (ou branches et feuilles),
- diagrammes,
- boîtes à pattes

et quatre étapes :

- validation, listage des données,
- regroupement, graphiques des distributions observées et analyses,
- calcul des paramètres,
- construction des boîtes à pattes.

La description est faite série par série (cuve par cuve), en prenant soin cependant de choisir et de positionner les échelles de façon à pouvoir comparer, par la suite, les représentations graphiques des différentes séries.

1 - Étape 1 : validation et listage des données

a) - Validation

Il s'agit de détecter les valeurs *suspectes*, c'est-à-dire les éventuelles erreurs de saisie, ainsi que les valeurs s'éloignant beaucoup du groupe ou des valeurs de référence de la variable dans les conditions étudiées. Ceci est réalisé en ordonnant les données, par exemple en utilisant les commandes de tri des tableurs (comme on l'a fait pour présenter les tableaux précédents), puis en inspectant le début et la fin des séries réordonnées. Dans la pratique, il faut cependant prendre soin de bien garder une trace de l'ordre initial. Lorsqu'on a des couples de variables, (lorsqu'on mesure deux caractères sur le même individu, par exemple), il faut veiller à ne pas désapparier les observations.

D'après la connaissance que nous avons des ordres de grandeur de la variable : *nombre de bactéries par millilitre*, dans les conditions d'étude, les séries du document 1 ne présentent pas de valeurs suspectes.

b) - Listage

Si les données ont été saisies dans un fichier informatique, il est important d'en sortir une trace papier, car on est plus à l'aise, par la suite, pour les consulter et les exploiter.

Si l'on a des fiches de saisie manuelle, il est toujours utile de faire un tableau récapitulatif des données, en identifiant bien variables, séries (ou groupes), individus. Il peut être intéressant d'avoir les séries triées, mais souvent l'ordre de saisie des valeurs est important et il faut toujours en garder une trace.

Dans cette étude on considère un caractère : le nombre de bactéries par millilitre. Chaque donnée statistique est constituée d'une mesure, prise à partir du tuyau de vidange.

2 - Étape 2 : regroupement des données, graphiques des distributions observées et analyse

C'est à cette étape que l'on met en évidence les distributions observées et qu'on les analyse en les illustrant.

a) - Regroupement des données en classes

Avant d'effectuer un regroupement des données d'un caractère quantitatif, il faut d'abord se fixer, soit un nombre de classes, soit une étendue de classes, selon l'utilisation ultérieure qui sera faite du regroupement obtenu : simple diagramme en bâtons, *tige et feuilles*, comparaison à des normes, etc.

Il existe des formules pour fixer, *a priori*, un nombre de classes[4], par exemple : $p = Ent(\sqrt{n})$, n étant l'effectif de la série. En fait il n'y a pas de recette miracle. Il est souvent intéressant de faire plusieurs découpages et d'en comparer les résultats.

Qu'en est-il pour les trois séries observées ?

Pour les séries de notre exemple, nous avons choisi de faire des classes de 20 en 20 car on obtient ainsi des limites de classes qui correspondent aux valeurs de référence des normes de qualité hygiénique du matériel. De 0 à 40 bactéries par ml, la qualité est dite *très bonne* ; de 40 à 80, la qualité est dite *bonne* ; de 80 à 120, la qualité est dite *médiocre* et il faut l'améliorer dans les 15 jours suivants ; de 120 à 200, la qualité est dite *mauvaise* et il faut procéder à une désinfection supplémentaire avant réutilisation des cuves ; au-delà de 200, il faut revoir complètement la procédure de désinfection et procéder à un lavage méticuleux des cuves infectées.

On peut ensuite faire le tableau des effectifs ou une représentation graphique, mais il est bien plus intéressant de réaliser les *tige et feuilles* (ou *branches et feuilles*) comme cela est présenté ci-dessous. (La méthode de construction d'un *tige et feuilles* est donnée dans l'annexe 1.)

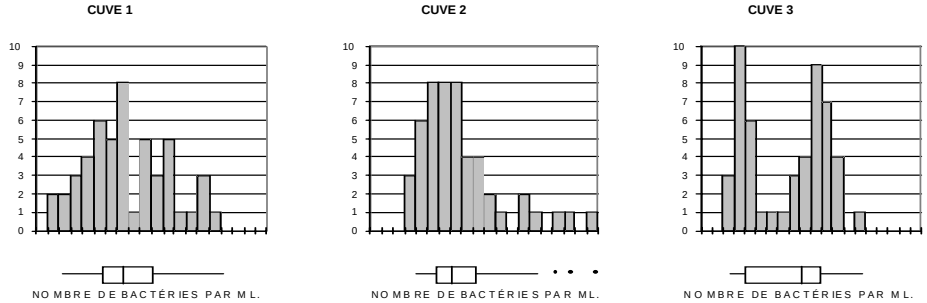
CUVE 1			CUVE 2			CUVE 3		
	0a			0a			0a	
2	0b	33		0b			0b	
4	0c	44		0c		3	0c	455
7	0d	667	3	0d	777	13	0d	6677777777
11	0e	8899	9	0e	888999	19	0e	888999
17	1a	000111	17	1a	00001111	20	1a	0
22	1b	22333	(25)	1b	22223333	21	1b	2
(8)	1c	44444455	(25)	1c	44455555	22	1c	5
20	1d	7	17	1d	6777	(25)	1d	667
19	1e	89999	13	1e	8889	(25)	1e	8899
14	2a	000	9	2a	11	21	2a	000001111
11	2b	23333	7	2b	2	12	2b	2222333
6	2c	5	6	2c		5	2c	4445
5	2d	7	6	2d	66	1	2d	
4	2e	889	4	2e	8	1	2e	9
1	3a	0		3a			3a	
				HI	**32, 35, 38**			

Document 2 : *Tige et feuilles* de la qualité du nettoyage mesurée par le nombre de bactéries totales par ml dans l'eau du dernier rinçage

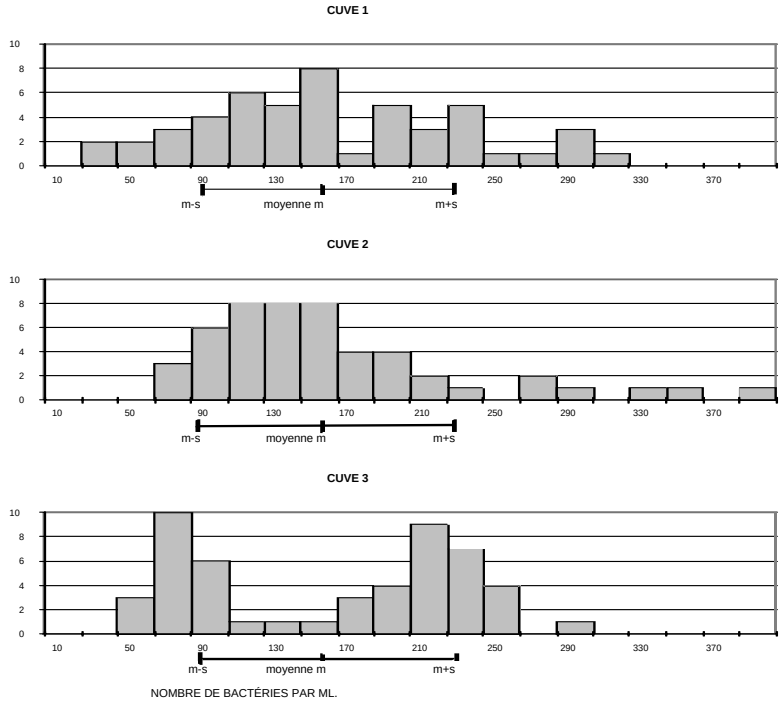
[4] On pourra consulter HOAGLIN et al. (1983) pour la justification de certaines formules classiques.

b) - Graphiques des distributions observées série par série

Le *tige et feuilles* est une représentation semi-graphique qui permet en même temps de regrouper les données en classes et d'en visualiser la distribution observée, comme un diagramme horizontal. Sur le document 3, nous faisons figurer les diagrammes des effectifs et les boîtes à pattes des trois séries et sur le document 4 qui suit, uniquement les trois diagrammes. Cette autre présentation ne fait-elle pas double emploi ? On peut y réfléchir, un élément de réponse sera apporté au paragraphe 5 suivant.



Document 3 : Diagrammes des effectifs et boîtes à pattes



Document 4 : Diagrammes des effectifs

c) - Analyse des distributions observées

Nous proposons de nous intéresser aux quatre critères les plus couramment utilisés, pour guider l'analyse exploratoire de chacune des trois séries (les trois distributions) :

- recherche des structures,
- analyse de la symétrie,
- comparaison des dispersions,
- comparaison des positions.

(i) Structure de chaque distribution observée

La question que l'on doit d'abord se poser est celle de savoir si l'on est bien en présence d'une série homogène, c'est-à-dire issue d'une seule population. N'est-on pas en présence d'un mélange de distributions, signe d'un mélange de populations ?

Un tel mélange se manifeste, en général, par des distributions à plusieurs *pics*, appelés aussi les *modes*.

Qu'en est-il pour les 3 séries observées ?

En analysant les diagrammes du document 4, on a une suspicion sur la distribution de la série de la cuve 1, à laquelle nous n'accorderons qu'une importance relative, vu le caractère peu marqué des deux modes et la faiblesse de l'effectif. Les méthodes employées pour séparer les populations s'utilisent, dans la pratique, avec des effectifs de séries au moins égaux à 200.

Il n'y a rien à dire sur la distribution de la série de la cuve 2.

Quant à la distribution de la série de la cuve 3, elle est très nettement bimodale. Une étude rigoureuse consisterait, ici, à tenter d'identifier et/ou de séparer les *sous-populations*, pour reprendre le travail descriptif sur chacune des séries du mélange. L'origine de ce mélange peut être recherché, entre autres, dans les conditions opératoires dans lesquelles ont été faites les mesures. Si, par exemple, elles ont été réalisées à deux périodes séparées par une pause d'une heure, c'est un temps suffisant pour donner lieu à une importante multiplication microbienne en cas de désinfection insuffisante. Ces informations doivent (normalement) figurer dans le rapport d'exécution du contrôle, faute de quoi il est beaucoup plus difficile de justifier la séparation des populations ou d'en faire l'identification.

Ne possédant pas plus d'information sur les conditions de réalisation des mesures, nous n'avons pas séparé les deux populations.

Les descriptions effectuées par la suite, sur la série de la cuve 3, sont donc un pur exercice d'école, montrant ce qu'il ne faut pas faire !

Signalons enfin, qu'il existe des méthodes numériques et/ou graphiques, de séparation de populations gaussiennes en mélange.

(ii) Symétrie de chaque distribution observée

En analyse exploratoire, la symétrie d'une distribution s'évalue à l'oeil, en observant le *tige et feuilles* ou le diagramme des effectifs et en cherchant s'il existe un axe de symétrie. Rappelons que dans une analyse exploratoire, on tente de voir rapidement et simplement certaines caractéristiques des distributions observées.

La symétrie d'une distribution s'évalue série par série, car c'est un critère qui peut s'apprécier dans l'absolu : on peut dire, simplement en observant le *tige et feuilles* d'une distribution, si elle est à peu près symétrique ou non.

Qu'en est-il pour les trois séries observées ?

La distribution de la série de la cuve 1 (cf. le document 3) présente une dissymétrie gauche peu marquée, c'est-à-dire avec des effectifs plus importants vers les valeurs faibles du caractère. On dit aussi dissymétrie avec le poids principal à gauche. En première approximation, on pourra l'assimiler à une distribution symétrique, toujours eu égard à la faiblesse de l'effectif de la série. Une dissymétrie de la même ampleur, avec un effectif plus important, nous aurait sans doute amenés à approfondir la question. Rappelons encore que nous sommes en analyse exploratoire et qu'il s'agit seulement de glaner des impressions, sans chercher à conjecturer sur des distributions théoriques.

La distribution de la série de la cuve 2 présente une forte dissymétrie gauche. Dans ce cas, si l'on a besoin de la comparer à un modèle de distribution théorique, on ne pourra utiliser ni le modèle gaussien (illustré par la courbe en cloche), ni un autre modèle symétrique.

De plus, dans ce type de distribution statistique où l'on ne peut supposer l'indépendance des données, l'interprétation pratique de certains paramètres tels que la moyenne et l'écart-type est difficile et de peu d'intérêt, voire impossible, comme nous le verrons à l'étape 3 suivante, contrairement aux situations où les données peuvent être considérées comme constituant un échantillon résultant de prélèvements aléatoires indépendants.

(iii) Dispersion des trois distributions observées

Contrairement aux deux critères précédents, dans la pratique, l'évaluation de la dispersion d'une distribution ne se fait pas dans l'absolu. Elle n'a d'intérêt que lorsqu'on la compare à la dispersion d'une distribution de référence (un modèle théorique ou un étalon), par exemple une gaussienne dont on superpose le graphique de la distribution théorique à celui de la distribution observée, ou bien lorsque l'on compare les dispersions de plusieurs distributions.

En analyse exploratoire, on compare d'abord les dispersions des distributions à l'oeil, sur les graphiques (*tige et feuilles*, diagrammes ou boîtes à pattes),

simplement pour détecter de grosses différences entre elles. On pourra ensuite affiner l'analyse en calculant des paramètres, dont le plus utilisé est l'écart-type.

Qu'en est-il pour les trois séries observées ?

Nous n'avons pas, dans le cas concret traité, de distribution de référence, on se contentera donc de comparer les distributions entre elles. Elles ne présentent pas visuellement d'écart important de dispersion. La distribution de la série de la cuve 2 se distingue de suite par ses trois valeurs extrêmes, notées HI sur le *tige et feuilles*. On trouvera dans les annexes 1 et 2 comment se caractérisent ces valeurs. L'expérience du biologiste nous dit qu'il n'y a pas lieu de s'alarmer de ces quelques valeurs s'éloignant du groupe.

Anticipons un peu sur le paragraphe 3 (calcul des valeurs des paramètres et analyse) : les trois distributions ont des écarts-types ne différant qu'à partir de la troisième décimale. Ce paramètre est donc insuffisant pour rendre compte de la différence entre les trois distributions observées que l'on constate bien visuellement sur les *tige et feuilles*.

On peut utiliser l'écart inter-quartile[5], $Q_3 - Q_1$, qui permet de mieux se rendre compte d'une différence entre les trois distributions. On trouvera ces valeurs dans le document 5, objet du paragraphe 3.

Mais d'une façon générale, bien plus qu'une petite différence entre les paramètres de dispersion, c'est bien les différences fondamentales dans la *forme* des distributions observées qu'il importe de *voir*.

(iv) Position des trois distributions observées

Comme dans le cas de la dispersion, dans la pratique, l'évaluation de la position d'une distribution ne se fait pas dans l'absolu. On la compare à une valeur de référence, ou on compare les positions respectives des distributions des groupes (ou séries) que l'on étudie.

Respectant l'*esprit* de l'analyse exploratoire, c'est d'abord à l'oeil que l'on compare, sur les graphiques, les positions des différentes distributions. On peut ensuite affiner par le calcul des paramètres de position.

Qu'en est-il pour les trois séries observées ?

Il n'y a pas, visuellement, de grosses différences de position entre les trois distributions. Anticipons, là encore, sur le paragraphe 3 (calcul et analyse des paramètres) : les moyennes des trois séries sont très proches, ne *révélant* pas les différences entre les distributions.

---

[5] Pour les définitions, on pourra se reporter à l'article de J. C. GIRARD *Quartiles, déciles et tutti quantiles* dans ce même volume.

Examinons un autre paramètre de position : la médiane. Elle différencie bien les trois distributions observées et s'interprète facilement : au moins 50 % des observations sont inférieures ou égales à la médiane et au moins 50 % sont supérieures ou égales à la médiane. On trouvera ces valeurs dans le document 5 dont le mode d'emploi figure au paragraphe suivant.

Rappelons que dans le cas de la série 3, les calculs de paramètres n'ont pas beaucoup de sens, puisqu'il y a suspicion de mélange de populations. Nous dirons, ici aussi, que l'important n'est pas une différence entre les valeurs de certains paramètres, mais bien plus, les différences entre les distributions, dont les paramètres peuvent ne pas rendre compte.

Nous avons vu que moyenne et écart-type ne différencient pas du tout les trois distributions étudiées. Il est donc important d'utiliser des paramètres variés et surtout de se dire que deux (voire même trois) paramètres ne fourniront en général qu'une piètre description.

3 - Étape 3 : calcul des valeurs des paramètres des résumés en chiffres

L'objectif de cette étape est de fournir les valeurs de paramètres permettant de quantifier certaines des différences observées précédemment entre les distributions. Ces valeurs permettront aussi, à l'étape suivante, de construire un résumé graphique de chaque série, appelé suivant les auteurs : diagramme en boîte, boîte à pattes, boîte à moustaches ou boîte de dispersion.

Les principaux paramètres utilisés pour l'analyse exploratoire d'une série d'un caractère quantitatif sont les suivants : la médiane (Q_2), les quartiles (Q_1 et Q_3)[6], le minimum et le maximum de la série. Ils sont faciles à déterminer et ne sont sensibles qu'à l'ordre des valeurs et non aux valeurs elles mêmes.

A ces paramètres viennent s'ajouter d'autres valeurs : $H = 1{,}5\,(Q_3 - Q_1)$, $Linf = Q_1 - H$, $Lsup = Q_3 + H$, et les bornes Binf égale à la plus petite valeur de la série supérieure à Linf, et Bsup égale à la plus grande valeur de la série inférieure à Lsup.

Pour faciliter la lecture du document 5, nous présentons ci-dessous, un mode d'emploi détaillé de ce que l'on appelle le *résumé en chiffres* ou *boîte à chiffres*.

---

[6] cf. l'article précédent de J. C. GIRARD. Nous adoptons ici les définitions suivantes : dans la série ordonnée, la médiane est la valeur de « rang fictif » $\frac{n+1}{2}$ (si n est pair, c'est la demi somme des données de rangs $\frac{n}{2}$ et $\frac{n}{2}+1$), le quartile Q_1 est la valeur (éventuellement interpolée) de « rang fictif » $\frac{1}{2}\,\mathrm{Ent}\left(\frac{n+1}{2}+1\right)$ et Q_3 la valeur (éventuellement interpolée) de « rang fictif » $n+1-\frac{1}{2}\,\mathrm{Ent}\left(\frac{n+1}{2}+1\right)$.

Les valeurs des principaux paramètres utiles en analyse exploratoire y sont rappelées.

Identification de la série	# n : effectif de la série	
médiane	Q_2	
quartiles inférieur et supérieur :	Q_1	Q_3
minimum et maximum	Min	Max

On peut compléter ce résumé par des éléments servant à la construction des boîtes à pattes (voir le paragraphe 4) et à approfondir les analyses :

H (sert au calcul de Linf et Lsup)	$H = 1,5\ (Q_3 - Q_1)$	
Linf et Lsup (servent au calcul des bornes)	$Linf = Q_1 - H$	$Lsup = Q_3 + H$
Bornes : inférieure et supérieure (servent dans la construction des boîtes à pattes)	B. inf. est la plus petite valeur de la série supérieure à Linf.	B. sup. est la plus grande valeur de la série inférieure à Lsup.
(approfondissement de l'analyse et de la comparaison des distributions)	Proportion (en %) des valeurs comprises entre B. inf. et B. sup. m : moyenne arithmétique de la série s : écart-type de la série Proportion (en %) des valeurs comprises entre m – s et m + s Proportion (en %) des valeurs comprises entre m – 2s et m + 2s	

On trouvera dans l'annexe 2 une autre application numérique guidée, pour se familiariser avec ces outils particuliers, dont il n'est pas inutile de préciser qu'ils sont maintenant disponibles sur beaucoup de logiciels statistiques graphiques modernes (STATGRAPHICS, SYSTAT, MINITAB, SPSS, SPAD, S, R[7]...).

Qu'en est-il pour les trois séries observées ?

On obtient pour les trois cuves les mêmes moyennes m = 157,24 et les mêmes écarts-types s = 70,52. Nous pouvons compléter l'analyse et la comparaison des graphiques des trois distributions observées, à l'aide des résumés en chiffres présentés dans le document 5 .

[7] R est téléchargeable gratuitement sur http://www.r-project.org (A Programming Environment for Data Analysis and Graphics, 1999-2004, R Development Core Team, from the R-project).

Cuve 1	# 50	
médiane	145,291	
quartiles	106,215	203,470
Minimum et maximum	34,060	307,288
H	145,883	
Linf ; Lsup	– 39,668	349,323
B.inf. ; B.sup.	34,1	307,3
% entre les bornes	$\frac{50}{50}$ = 100 %	
% dans [m – s ; m + s]	$\frac{33}{50}$ = 66 %	
% dans [m – 2s ; m + 2s]	$\frac{49}{50}$ = 98 %	

Cuve 2	# 50	
médiane	138,750	
quartiles	107,112	180,003
Minimum et maximum	72,270	388,430
H	109,137	
Linf ; Lsup	– 2,225	289,340
B.inf. ; B.sup.	72,3	289,3
% entre les bornes	$\frac{47}{50}$ = 94 %	
% dans [m – s ; m + s]	$\frac{38}{50}$ = 76 %	
% dans [m – 2s ; m + 2s]	$\frac{48}{50}$ = 96 %	

Cuve 3	# 50	
médiane	178,523	
quartiles	78,889	215,186
Minimum et maximum	44,454	292,867
H	204,445	
Linf ; Lsup	- 125,560	419,631
B.inf. ; B.sup.	44,5	292,9
% entre les bornes	$\frac{50}{50}$ = 100 %	
% dans [m – s ; m + s]	$\frac{24}{50}$ = 48 %	
% dans [m – 2s ; m + 2s]	$\frac{50}{50}$ = 100 %	

Document 5 : Résumés en chiffres

Ces résumés, que l'on peut présenter judicieusement sur une simple colonne, indiquent déjà bien les principales différences entre les trois distributions, déjà vues lors de l'analyse des *tige et feuilles*. La représentation graphique de ces résumés en chiffres est faite par les boîtes à pattes que nous abordons au paragraphe 4.

Par exemple un écart important entre moyenne et médiane est un indice de dissymétrie. Les proportions (en %) des valeurs comprises entre $m - s$ et $m + s$, puis $m - 2s$ et $m + 2s$, comparées à 68 % et à 95 %, qui sont les valeurs théoriques fournies par la distribution de Gauss (conforme à la loi normale centrée réduite, illustrée par la courbe en cloche), donnent des indices d'écart à la normalité. Mais il faut se méfier de l'utilisation abusive des valeurs des paramètres du modèle gaussien.

Les valeurs lues directement dans le document 5, nous indiquent des écarts notables entre les distributions des séries 1 et 2, des séries 1 et 3 et des séries 2 et 3. Nous sommes bien *avertis* que les distributions des séries diffèrent entre elles, que la série 1 se rapproche d'un modèle gaussien alors que les séries 2 et 3 s'en éloignent.

Il est intéressant de comparer les valeurs théoriques, avec les valeurs observées, indiquées sur le document 5 : par exemple pour la série 2, la moyenne vaut 157,24 , valeur éloignée de la médiane qui vaut 138,75. La proportion, en %, des valeurs comprises entre $m - s$ et $m + s$ vaut 76 %, à comparer aux 68 % du modèle gaussien.

Seule la distribution de la série de la cuve 1 ne présente pas d'écart important avec les valeurs théoriques du modèle gaussien.

Le choix d'un modèle de distribution théorique représentant correctement la distribution observée, n'est pas chose facile, surtout lorsqu'on s'éloigne du modèle gaussien. C'est une des raisons pour lesquelles on essaie souvent de s'en rapprocher.

* Dans la pratique, lorsque l'on est en présence d'une distribution observée dissymétrique et que la connaissance que l'on a du phénomène étudié ne nous permet pas de choisir un modèle théorique précis, on essaie de trouver un changement de variable qui rende la distribution observée à peu près symétrique et aussi moins dispersée.

Parfois, lorsque que l'on a obtenu ces deux effets, la distribution se rapproche d'une gaussienne. Les effets des transformations que l'on essaie, s'apprécient, dans un premier temps (exploration oblige), avec tige et feuilles ou diagrammes.

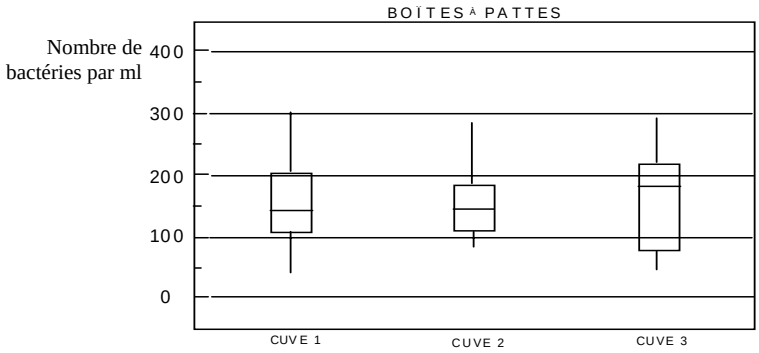
4 - Étape 4 : construction des boîtes à pattes

Les détails techniques de la construction des boîtes à pattes sont donnés dans les références citées plus haut et présentés dans l'annexe 2 de cet article ainsi que dans l'article de Jean Claude GIRARD (1993). Tous les éléments nécessaires à leur construction figurent dans les résumés en chiffres.

Lorsque ces graphiques sont réalisés à la main, il est intéressant de repérer sur la liste des données triées, les bornes et les valeurs situées à l'extérieur des bornes. Le tracé des graphiques en est grandement facilité.

Qu'en est-il pour les trois séries observées ?

Le document 6 représente les boîtes à pattes des trois séries.



Document 6 : Boîtes à pattes représentant la qualité du nettoyage mesurée par le nombre de bactéries totales par ml dans l'eau du dernier rinçage

L'analyse des boîtes à pattes permet de retrouver la plupart des caractéristiques mentionnées à l'étape 2, dans l'analyse des distributions.

On y remarque facilement les caractères de dissymétrie de chacune des distributions, on y compare rapidement les dispersions et les positions des trois distributions.

Rappelons que si une distribution est symétrique, la moyenne est égale à la médiane et celle-ci se trouve au milieu des quartiles et que la réciproque est fausse. Mais, dans la pratique, une boîte symétrique est un bon indicateur de la symétrie de la distribution, que l'on peut confirmer en analysant le *tige et feuilles*.

Le seul élément n'apparaissant pas sur les boîtes à pattes est la structure de la distribution de la cuve 3. Elle n'est pas non plus visible simplement à partir des résumés en chiffres. N'omettons donc jamais le passage par le *tige et feuilles* ou par le diagramme des effectifs (ou des fréquences).

5 - Éléments complémentaires

Il n'est pas inutile de rappeler ici l'utilisation du *tige et feuilles*, bien que dans notre méthodologie, il apparaisse avant et que l'on est donc supposé être en possession des informations qu'il nous apporte**.** L'analyse exploratoire se pratiquant à l'aide de l'outil informatique, les contraintes imposées par beaucoup de logiciels font qu'il est beaucoup plus facile et rapide de *sortir* les boîtes à pattes, ce qui est donc, souvent, la première chose faite.

On a alors trop tendance à oublier les éléments des étapes précédentes, suite sans doute, au même sentiment de satisfaction que lorsque, il y a 25 ans, à l'apparition des premières calculatrices avec fonctions statistiques, on avait calculé une moyenne voire même, comble du raffinement, un écart-type !

La présentation sur une page, de certains types de graphiques, ne se fait pas au hasard. Elle obéit à des règles bien précises, liées à la façon dont on perçoit ces

objets graphiques. Ainsi notre œil est habitué à la lecture horizontale. Il sera donc bien plus efficace pour détecter des différences entre graphiques si ceux ci sont orientés et juxtaposés horizontalement, comme les *tige et feuilles* du document 2. Sur le document 3 les diagrammes mis côte à côte ne sont pas présentés correctement et sont donc très difficilement comparables. Le document 4 les présente donc disposés verticalement, ce qui permet de mieux comparer les distributions représentées, mais cette direction de lecture n'est pas confortable car elle n'exploite pas les caractéristiques particulières de notre vision.

Rappelons aussi, par exemple, que ce que les tableurs standard appellent histogramme n'en sont aucunement, au sens statistique du terme. Ce sont simplement des diagrammes en barres. Les tableurs ne sont pas conçus pour le traitement statistique des données, même si les nouveaux standards intègrent de plus en plus de fonctions statistiques. On n'a encore rien trouvé de plus efficace que même une simple chaîne de traitement statistique (par exemple : MINITAB ou R), pour traiter des données. Les tableurs doivent rester dans la fonction dans laquelle ils excellent, la gestion des tableaux de données.

III - Conclusion

Résumons d'abord les quatre étapes de la description d'une série univariée à une variable quantitative :

Étape 1 : Validation et listage des données

Validation pour repérer et corriger les erreurs

Listage pour avoir un support de travail

Étape 2 : Regroupement des données, graphiques des distributions observées et analyse

Regroupement pour réaliser un découpage en classes des valeurs de la série

Graphique pour réaliser un *tige et feuilles* ou un diagramme

Analyse des distributions observées

- Structure de chaque distribution : est-on bien en présence d'une seule population ?
- Symétrie de chaque distribution, éventuellement diagnostic d'un changement de variable
- Comparaison des dispersions des distributions, par rapport à des références ou entre elles
- Comparaison des positions des distributions, par rapport à des références ou entre elles

Étape 3 : Calcul des valeurs des paramètres des résumés en chiffres

Support chiffré à l'analyse des distributions

Étape 4 : Construction des boîtes à pattes

Résumé graphique efficace des distributions à étudier.

On est à des années lumières du simple calcul moyenne, écart-type, que l'on voit encore trop souvent fleurir dans les rapports *techniques* en tout genre, en guise de statistique descriptive.

La statistique descriptive c'est la description et l'analyse des distributions observées, en suivant une méthodologie précise qui permet d'éviter les principaux pièges. C'est aussi et de façon non moins importante, gérer des données c'est-à-dire les préparer et les structurer pour qu'elles puissent être utilisées par les outils logiciels.

Avec l'avènement des logiciels graphiques puissants et bon marché, l'analyse exploratoire et les méthodes développées par TUKEY sont les outils privilégiés de cette description. Ils deviennent aussi le passage incontournable d'une formation pratique, réaliste, opératoire à la statistique descriptive moderne.

Références bibliographiques

CHAMBERS, J. M., CLEVELAND, W. S., KLEINER, B., TUKEY, P. A, *Graphical Methods For Data Analysis*, Wadworth & Brooks Cole, Pacific Grove, 395 pp., 1983

GIRARD, J. C. , La médiane pour quoi faire ? Un exemple d'utilisation : les boîtes de dispersion, *Colloque de la commission Inter-IREM Statistique et Probabilités*, Toulouse, 14-15 mai 1993.

HOAGLIN, D., MOSTELLER, F., TUKEY, P. A., *Understanding Robust And Exploratory Data Analysis*, Wiley & Sons, 1983.

TUKEY, J. W., *Exploratory Data Analysis*, Addison Wesley, 5th edition, 499 pp., 1977.

WONNACOTT, T. H., WONNACOTT, R. J., *Introductory Satistics*, John Wiley & Sons, 711 pp., 1990.

ANNEXE 1 : Construction d'un graphique *tige et feuilles* (stem and leaf - TUKEY 1977)

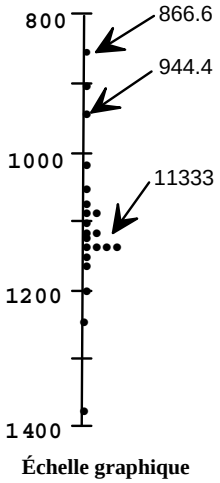
Il s'agit d'étudier la distribution de la variable *Gain Moyen Quotidien* de poids (GMQ) chez des jeunes bovins (taurillons) de race charolaise, issus de saillie naturelle (pour les comparer par la suite à ceux issus de l'insémination artificielle). On utilise les mesures effectuées sur un échantillon aléatoire et simple de 20 animaux. L'unité utilisée est le gramme par jour (gr/j). On obtient la série x_i que l'on peut décrire des façons suivantes :

Représentations numériques non résumées : les *listes*

x_i dans l'ordre de saisie	$x_{(i)}$ par ordre croissant	$x^{(i)}$ par ordre décroissant
$x_1 = 900,0$	$x_{(1)} = 866,6$	$x^{(1)} = 1\,377,7$
$x_2 = 944,4$	$x_{(2)} = 900,0$	$x^{(2)} = 1\,255,5$
$x_3 = 1\,066,6$	$x_{(3)} = 944,4$	$x^{(3)} = 1\,155,5$
$x_4 = 1\,111,1$	$x_{(4)} = 1\,011,1$	$x^{(4)} = 1\,144,4$
$x_5 = 1\,255,5$	$x_{(5)} = 1\,055,5$	$x^{(5)} = 1\,133,3$
$x_6 = 1\,155,5$	$x_{(6)} = 1\,066,6$	$x^{(6)} = 1\,133,3$
$x_7 = 1\,133,3$	$x_{(7)} = 1\,077,7$	$x^{(7)} = 1\,133,3$
$x_8 = 1\,133,3$	$x_{(8)} = 1\,077,7$	$x^{(8)} = 1\,133,3$
$x_9 = 1\,011,1$	$x_{(9)} = 1\,088,8$	$x^{(9)} = 1\,111,1$
$x_{10} = 1\,111,1$	$x_{(10)} = 1\,100,0$	$x^{(10)} = 1\,111,1$
$x_{11} = 1\,055,5$	$x_{(11)} = 1\,111,1$	$x^{(11)} = 1\,100,0$
$x_{12} = 1\,133,3$	$x_{(12)} = 1\,111,1$	$x^{(12)} = 1\,088,8$
$x_{13} = 1\,133,3$	$x_{(13)} = 1\,133,3$	$x^{(13)} = 1\,077,7$
$x_{14} = 1\,144,4$	$x_{(14)} = 1\,133,3$	$x^{(14)} = 1\,077,7$
$x_{15} = 1\,100,0$	$x_{(15)} = 1\,133,3$	$x^{(15)} = 1\,066,6$
$x_{16} = 1\,377,7$	$x_{(16)} = 1\,133,3$	$x^{(16)} = 1\,055,5$
$x_{17} = 1\,077,7$	$x_{(17)} = 1\,144,4$	$x^{(17)} = 1\,011,1$
$x_{18} = 1\,077,7$	$x_{(18)} = 1\,155,5$	$x^{(18)} = 944,4$
$x_{19} = 1\,088,8$	$x_{(19)} = 1\,255,5$	$x^{(19)} = 900,0$
$x_{20} = 866,6$	$x_{(20)} = 1\,377,7$	$x^{(20)} = 866,6$

On a la relation : $x_{(i)} = x^{(n-i+1)}$

Représentation graphique non résumée la plus simple : les points sur un axe gradué



Échelle graphique

La valeur de la mesure effectuée sur chaque individu est repérée sur la droite par •

Construction d'un *tige et feuilles* :

- Le *tige et feuilles* se construit en regroupant les données en classes. Il faut donc se fixer, *a priori,* soit un nombre de classes, soit une étendue de classes. Il n'y a pas de nombre de classes idéal.
- Les valeurs appartenant à une classe sont représentées sur une ligne appelée *tige*, comprenant l'insertion, à gauche de la ligne verticale et les *feuilles*, à droite, qui sont les mesures faites sur les individus de la classe. En général, une valeur est représentée par un chiffre (une feuille), le chiffre des dixièmes, des unités, des dizaines, des centaines etc. selon l'unité choisie.
- Les classes utilisées déterminent les caractères - chiffres et symboles - figurant à l'insertion ; tandis que l'unité permet d'indiquer les valeurs que représentent les feuilles.
- La tige se positionne sur une échelle semi-graphique verticale, déterminée par la hauteur des lignes et interlignes ; alors que les feuilles sont juxtaposées sur une échelle semi-graphique horizontale, déterminée par la largeur des caractères (chiffres).

Représentation :

Le *tige et feuilles* illustre éléments et les caractéristiques suivants d'une variable, dans un échantillon ou une population finie si c'est une population qui est représentée :

- la distribution observée de la variable étudiée, par le profil engendré par l'extrémité des branches. Les principales caractéristiques observables d'une distribution, sont la symétrie, la dispersion et la position, à condition d'avoir un ensemble homogène**,** c'est-à-dire pas un mélange de populations, donc pas un mélange de distributions ;
- la branche médiane qui est toujours identifiée dans la colonne des effectifs et les valeurs éloignées qui peuvent aussi être identifiées après calcul de certaines statistiques de rang ;
- chaque individu de la série, qui est représenté et identifié par chaque feuille.

Utilisation :

Cette représentation permet :

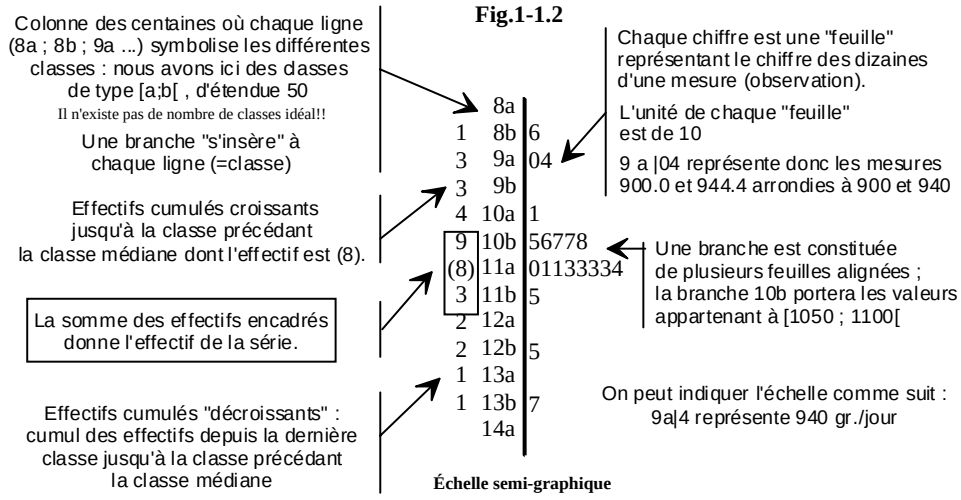
- de valider les données, en repérant et en identifiant les valeurs éloignées, ce qui permet de corriger les erreurs ;
- de détecter les structures (par exemple, mélange de population), en repérant sur le profil engendré par l'extrémité des branches, la présence de plusieurs *pics* (les modes) ;
- de caractériser la distribution de la variable dans l'échantillon (ou dans la population si c'est la population qui est représentée), en analysant le profil

engendré par l'extrémité des branches ; les structures, les valeurs éloignées et la symétrie peuvent être *appréciées* dans l'absolu, c'est-à-dire à partir d'une seule série, alors que position et dispersion nécessitent, pour être commentées, d'être comparées soit à une référence, soit aux caractéristiques d'une autre série ;

- d'avoir une idée de la distribution de la variable dans la population dont a été tiré l'échantillon ;
- de choisir une transformation de variable et d'en observer l'effet ;
- de comparer plusieurs échantillons (ou groupes), en juxtaposant les représentations.

Graphique *tige et feuilles* pour l'exemple donné :

Fig.1-1.2



Variantes de branches et feuilles :

Fig.1-1.3

unité des feuilles : 10
11| 3 représente 1130 gr./jour
(classes d'étendue 100)

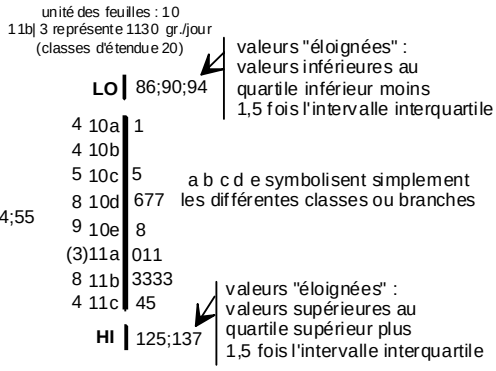
1 8 | 6
3 9 | 04
9 10 | 156778
(9) 11 | 011333345
2 12 | 5
1 13 | 7
14 |

Fig.1-1.4

unité des feuilles : 1
11| 33 représente 1133 gr./jour
(classes d'étendue 100)

1 8 | 66
3 9 | 00;44
9 10 | 11;55;66;77;77;88
(9) 11 | 00;11;11;33;33;33;33;44;55
2 12 | 55
1 13 | 77
14 |

Fig.1-1.5



ANNEXE 2 : Les boîtes à pattes ou boîtes de dispersion (boxplot - TUKEY 1977)

Reprenons l'étude de la variable GMQ (gr/j) et la série utilisée pour l'annexe 1.

Calcul de certaines statistiques de rang et présentation de la boîte à chiffres :

$x_{(1)} = 866,6$
$x_{(2)} = 900,0$
$x_{(3)} = 944,4$
$x_{(4)} = 1\,011,1$
$x_{(5)} = 1\,055,5$
$x_{(6)} = 1\,066,6$
$x_{(7)} = 1\,077,7$
$x_{(8)} = 1\,077,7$
$x_{(9)} = 1\,088,8$
$x_{(10)} = 1\,100,0$
$x_{(11)} = 1\,111,1$
$x_{(12)} = 1\,111,1$
$x_{(13)} = 1\,133,3$
$x_{(14)} = 1\,133,3$
$x_{(15)} = 1\,133,3$
$x_{(16)} = 1\,133,3$
$x_{(17)} = 1\,144,4$
$x_{(18)} = 1\,155,5$
$x_{(19)} = 1\,255,5$
$x_{(20)} = 1\,377,7$

STATISTIQUES	RANGS (approchés)	PROFONDEURS	VALEURS
minimum	1	1	$x_{(1)} = 866,6$
quartile inférieur (Q_1)	$\frac{n+1}{4}$	$\frac{\text{Ent}\left(\frac{n+1}{2}+1\right)}{2}$	$x_{(5,5)} = \frac{x_{(5)} + x_{(6)}}{2} = 1\,061,1$
médiane (Q_2)	$\frac{n+1}{2}$	$\frac{n+1}{2}$	$x_{(10,5)} = \frac{x_{(10)} + x_{(11)}}{2} = 1\,105,6$
quartile supérieur (Q_3)	$\frac{3(n+1)}{4}$	$\frac{\text{Ent}\left(\frac{n+1}{2}+1\right)}{2}$	$x_{(5,5)} = \frac{x_{(5)} + x_{(6)}}{2} = 1\,133,3$
maximum	n	1	$x_{(1)} = 1\,377,7$

Ces statistiques peuvent être présentées dans une boîte à chiffres :

Effectif	# 20	"saillie naturelle"	
Profondeurs	M 10,5	1105,6	
	Q 5,5	1061,1	1133,3
	1	866,6	1377,7

M : Médiane ; Q : Quartiles

Les autres valeurs nécessaires à la construction de la boîte à pattes

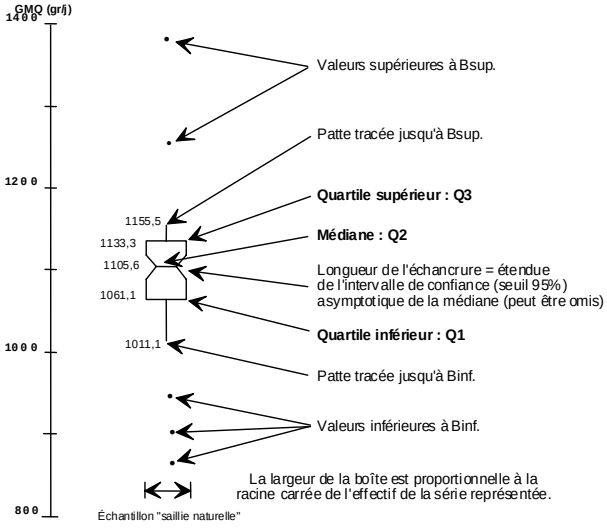
H=1,5*(Q3-Q1) = 108,3

Linf.=Q1-H = 952,8 Lsup.=Q3+H = 1241,7

Bsup. est la plus grande valeur de la série inférieure à Lsup. : 1155,5

Binf. est la plus petite valeur de la série supérieure à Linf. : 1011,1

Représentation graphique résumée : la boîte à pattes (il en existe quelques variantes)



Construction :

Conventionnellement, on utilise six valeurs permanentes pour représenter un échantillon :

- les deux valeurs Bsup. et Binf., délimitant les extrémités des pattes ;
- les premier (Q_1), deuxième (Q_2 ou médiane) et troisième quartiles (Q_3), délimitant les extrémités de la boîte ;
- l'effectif de la série, déterminant la largeur de la boîte.

La longueur de l'échancrure est celle de l'intervalle de confiance (approximation gaussienne) de la médiane, en général au seuil de confiance de 95 %. L'intervalle est : $\left[Q_2 - 1{,}57 \frac{R}{\sqrt{n}} ; Q_2 + 1{,}57 \frac{R}{\sqrt{n}} \right]$ où $R = Q_3 - Q_1$ et n est l'effectif de la série.

Les valeurs éloignées, c'est-à-dire supérieures à Bsup. ou inférieures à Binf., si elles existent, sont toutes marquées individuellement.

Représentation :

Ces diagrammes permettent de présenter les caractéristiques suivantes de la distribution de la variable dans l'échantillon (on l'appelle distribution observée) :

- valeur centrale, par la médiane ;
- autres valeurs de position, par les quartiles Q_1 et Q_3 ;
- dispersion, par la longueur de la boîte (écart interquartile) ;
- symétrie, par la position de la médiane entre les deux quartiles (condition nécessaire mais non suffisante) ;
- queues de distributions, par les pattes ;
- effectif, par la largeur de la boîte ;
- individus éloignés, par les valeurs marquées individuellement.

Robustesse :

Les boîtes étant construites à partir des paramètres basés sur les rangs, elles sont résistantes à l'introduction de valeurs extrêmes : il peut y en avoir jusqu'à 25 % sans altération de la forme de la boîte, ni de la longueur des pattes.

Utilisation :

Cette représentation résumée des données permet :

- de valider les données, les valeurs éloignées pouvant être rapidement détectées et identifiées ;
- d'avoir une idée sur certaines caractéristiques de la distribution de la variable dans l'échantillon, symétrie, dispersion, position ;

- d'estimer certains paramètres de population, si l'échantillon est aléatoire et simple ;
- de comparer plusieurs groupes ou échantillons : taille, homogénéité des dispersions, similitude des médianes et quartiles ;
- de choisir une transformation de variable et d'en observer les effets, par exemple pour réduire des écarts de dispersion trop importants entre échantillons ;
- une *exploration inférentielle*, en observant le degré de recouvrement des échancrures entre plusieurs boîtes, ce qui permet d'avoir une indication sur l'égalité des médianes des populations dont sont tirés les échantillons.

Il est intéressant de savoir que les quantiles d'un échantillon aléatoire et simple sont de bons estimateurs des quantiles de la population dont il est tiré.

Description d'une série statistique à deux variables quantitatives : modélisation non probabiliste par les méthodes d'ajustement[1]

Stéphan MANGANELLI

L'objectif de cet article est de donner une certaine vue d'ensemble sur un thème présent dans les programmes de mathématiques de nos classes (filière ES, Baccalauréat Professionnel, Brevet de Technicien, Brevet de Technicien Supérieur, …).

En sortant un peu des schémas stéréotypés classiques (le calcul du coefficient de corrélation r a priori pour donner le feu vert, une utilisation de la covariance un peu théorique, une confusion entre le modèle et la réalité…), j'essaie, imprégné de multiples débats et lectures diverses, de donner ma façon de voir et d'aborder cette partie du programme, si possible à travers une démarche pédagogique cohérente.

Avec des éclairages particuliers sur certains aspects pratiques, en mettant l'accent sur certains outils théoriques plutôt que d'autres et en proposant certaines ouvertures, j'espère que des collègues y trouveront un intérêt.

La plus grande partie des résultats théoriques n'est pas démontrée, mais peut se retrouver notamment dans les livres cités en bibliographie. Ce choix est volontaire, dans le sens où il ne me semble pas primordial dans un premier temps de décortiquer les outils théoriques mais au contraire de faire ressortir l'aspect pratique ; ce d'autant plus que l'utilisation des outils comme les calculatrices ou l'ordinateur apparaît incontournable et permet d'obtenir ces résultats (droite d'ajustement) ou de les vérifier (équation d'analyse de variance).

Par contre la machine ne fait pas l'analyse critique de la situation et de ses résultats : c'est aussi sur cette analyse exploratoire que j'ai voulu insister pour amener les élèves à réfléchir.

Enfin, j'ai essayé de donner la priorité à l'étude de problèmes *réels*, liés concrètement à des options (que je connais un peu) et d'éviter le plus possible (et ce n'est pas évident !) les exercices d'école.

Tout cela en essayant modestement d'être dans l'esprit des programmes, qui nous orientent de plus en plus vers « *un entraînement à la lecture active de l'information et à*

[1] Je cite en bibliographie les références qui m'ont permis d'échafauder cet article ; je tiens à remercier mon collègue Hubert RAYMONDAUD pour sa bienveillance permanente.

son traitement » ainsi que vers « *une initiation à la pratique d'une démarche scientifique globale* » avec comme objectif général afférent notamment à notre thème d'« *être capable de choisir et d'utiliser, dans une situation donnée, un modèle mathématique adapté au traitement de l'information présentée sous différent aspects* ».

Ce thème a donc pour objet l'étude de la corrélation entre deux caractères quantitatifs observés sur une même population ; il fait le lien entre le côté exploratoire de la statistique descriptive et l'aspect décisionnel de la statistique inférentielle.

Une première approche peut en être faite avec l'ajustement affine par la méthode des points moyens (comme en BTA par exemple), puis approfondie en Terminale ES puis en BTS, avec la méthode des moindres carrés, les changements de variables, etc.

I - Ajustement d'un nuage de points

1 - Exemple introductif

On souhaite répondre à la question suivante :

« Y a-t-il une corrélation entre le revenu mensuel d'un ménage et la somme d'argent que ce ménage a mis dans l'achat de sa dernière voiture ? »

Pour avoir des éléments de réponse, on sélectionne un échantillon de 10 ménages jugés représentatifs d'une certaine population et on relève, pour chacun d'eux, les informations regroupées dans le tableau suivant, où les salaires et les budgets sont exprimés en euros :

Ménages	1	2	3	4	5	6	7	8	9	10
Salaire mensuel	1 067	1 220	1 372	1 524	2 134	2 439	2 744	3 049	3 049	4 269
Budget voiture	5 350	7 600	5 350	4 550	13 700	12 200	15 250	18 300	21 350	22 900

Bien entendu les ménages interrogés ne le sont que comme représentants de l'ensemble de tous les ménages ; leur identité est secondaire, voire inutile, car le chercheur désire, à partir de ces informations, dégager une loi ou une règle reliant le budget d'un ménage *quelconque* et la dépense qu'il consacre à l'achat d'une voiture. Le but de l'étude est, dans ce cas, de trouver une relation entre les caractères statistiques, cette relation ne pouvant apparaître qu'au travers des ménages retenus dans l'analyse.

2 - Nuage de points

Différentes étapes sont nécessaires pour découvrir et expliciter cette relation, si toutefois elle existe.

La première étape – indispensable – consiste à *représenter graphiquement* les données. L'œil est un puissant outil d'analyse et de décision, capable de donner des informations que les indicateurs statistiques élémentaires ne permettent pas toujours d'obtenir.

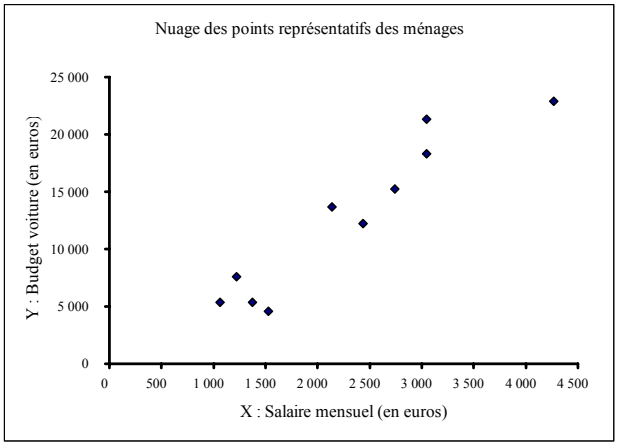
Dans la situation précédente, le graphique efficient est le nuage des points représentatifs des ménages. Pour réaliser cette représentation graphique :

- On trace, dans le plan, un système de deux axes orthogonaux tels que :
  - L'axe horizontal est gradué selon les valeurs du caractère *revenu*, noté X ;
  - L'axe vertical est gradué selon les valeurs du caractère *budget voiture*, noté Y.

  Dans la suite, on note $x = (x_1,\ldots, x_i,\ldots, x_n)$ la série statistique des valeurs observées du caractère X et $y = (y_1,\ldots, y_i,\ldots, y_n)$ la série statistique des valeurs observées du caractère Y. On notera $(x\,, y) = (x_i\,, y_i)_{1\leq i\leq n}$ la série double[2].
- On trace, pour chaque ménage présent dans l'analyse, un point ayant pour coordonnées $(x_i\,, y_i)$ les valeurs observées des deux caractères : son *revenu* en abscisse, son *budget voiture* en ordonnée (car on désire ici apprécier le budget en fonction du revenu et non le contraire comme ce pourrait être le cas pour un contrôle fiscal !). Ces points constituent le nuage noté N(X , Y).

On choisira les origines et les échelles des deux axes de telle sorte que tous les points ainsi marqués soient dans la zone réservée à ce graphique (sur une feuille ou à l'écran).

Pour l'exemple proposé, le nuage correspondant aux 10 ménages est :



3 - Exploitation du nuage

L'observation de ce nuage permet de répondre sommairement à l'interrogation relative à une relation entre le *revenu* et le *budget voiture* des ménages.

Ici, ce que l'œil voit immédiatement, c'est la forme de ce nuage : le nuage est étiré selon une certaine direction et les points semblent répartis de part et d'autre d'une droite fictive.

---

[2] Les valeurs observées x_i et y_i des caractères X et Y sont notées en italiques.

Ajuster *(de façon affine) le nuage, c'est lui associer une droite, de manière à ce que, en un sens à préciser mathématiquement, elle soit « le plus proche possible » de l'ensemble des points du nuage, reflétant au mieux sa position et son aspect étiré.*

Si on trace une telle droite et si on connaît le revenu d'un autre ménage (ne figurant pas parmi ceux qui ont servi à dessiner ce nuage), on peut se faire une idée sur son *budget voiture* : pour un ménage de revenu x, faute d'autres informations sur ses choix, on peut considérer que y est la valeur qu'il pense consacrer à l'achat d'une voiture, où y est l'ordonnée du point d'abscisse x sur la droite d'ajustement affine[3].

L'œil pourrait voir autre chose que l'étirement du nuage de points autour d'une droite : par exemple des points isolés, différents sous-groupes, relation autre qu'affine… Dans ce cas, on devra utiliser des traitements particuliers.

Pour l'instant, on se limite au cas où l'œil remarque que le nuage de points a bien une forme longiligne.

Remarque : Attention à l'effet d'échelle… !

4 - Droite d'ajustement

Une droite fictive ajustant au mieux le nuage de points a une équation réduite du type $y = a\,x + b$.

Pour tracer cette droite, diverses méthodes sont envisageables :

- Soit une méthode *manuelle* : on dessine avec précision le nuage des points et on trace à main levée (on a quand même droit à la règle !), une droite qui paraît bien *ajuster* l'ensemble des points.

  On peut alors, une fois la droite tracée, calculer les coefficients *a* et *b* de son équation à partir de la lecture des coordonnées de deux de ses points, ce qui conduit à deux équations linéaires aux inconnues *a* et *b*.

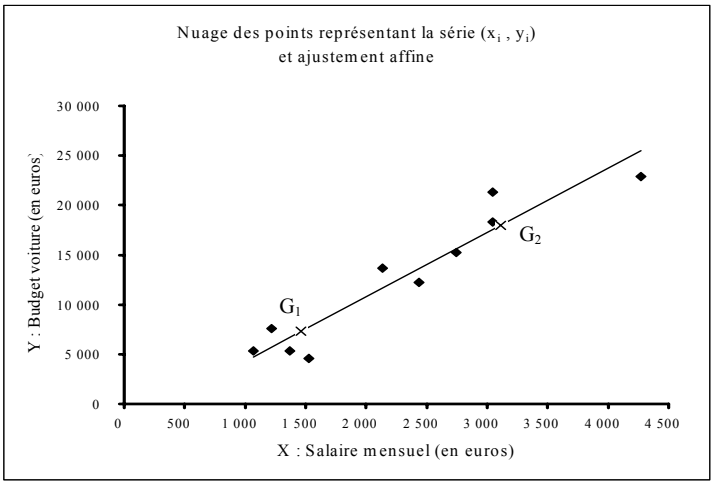
- Soit une méthode *objective* : on se donne un critère pour définir ce que l'on entend par *meilleure droite d'ajustement*, ce qui permet de calculer les paramètres *a* et *b* et de tracer la droite d'équation $y = a\,x + b$.

  Le critère généralement retenu est celui appelé *des moindres carrés*. C'est cette méthode qui apparaît dans divers programmes.

- Soit, comme on le verra par la suite, des méthodes adaptées à certaines données. Parmi celles-ci, la *méthode de Mayer* (proposée en particulier en BTA) : on partage les individus en deux groupes selon les valeurs x_i (en général celles qui sont inférieures à une borne donnée B et celles qui sont supérieures à B), ce qui constitue deux sous-nuages.

[3] Dans la suite, les lettres x, y (en caractères romains) désignent les coordonnées des points dans une représentation cartésienne.

On détermine les points moyens G_1 et G_2 pour chacun de ces sous-nuages et la droite d'ajustement est la droite (G_1G_2).



Dans notre exemple, avec B = 2 300, séparant le nuage en deux groupes de cinq points, on trouve G_1 (1 463 ; 7 310) et G_2 (3 110 ; 18 000).

L'équation réduite de (G_1G_2) est alors : $y = 6{,}49\,x - 2\,184{,}87$.

Par exemple, on peut estimer le *budget voiture* à environ 9 500 € pour un ménage ayant un salaire mensuel de 18 000 €.

Que peut-on prévoir pour un ménage dont le salaire est de 5 000 € ? Problème d'extrapolation, cf. ci-dessous).

5 - Une autre activité d'approche

L'étude porte sur la relation entre le poids et la taille des enfants à leur naissance. On a relevé le poids et la taille de 10 nouveau-nés (numérotés de 1 à 10 par ordre croissant de poids).

Enfant	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
Poids en kg	*2,4*	*2,6*	*2,7*	*3,0*	*3,2*	*3,3*	*3,5*	*3,6*	*3,8*	*4*
Taille en cm	*45*	*47*	*48*	*50*	*51*	*52*	*53*	*54*	*54*	*56*

1) - *Dans un repère orthogonal, placez les points ayant pour abscisse x le poids d'un enfant et pour ordonnée y la taille de cet enfant.*

2) - *Vous obtenez un nuage de points qui s'alignent plus ou moins sur une droite. L'étude qui suit propose une méthode permettant de déterminer cette droite.*

a) - *Partager l'ensemble des points constituant le nuage en deux sous-ensembles : celui des points correspondant aux enfants de 1 à 5 ; celui des points correspondant aux enfants de 6 à 10.*

b) - *Déterminer les coordonnées du point moyen de chaque sous-ensemble (l'abscisse du point moyen est égale à la moyenne des abscisses de chaque point du sous-ensemble, son ordonnée à la moyenne des ordonnées).*

Représenter la droite d'ajustement passant par les deux points ainsi déterminés.

c) - *Déterminer une équation de la droite d'ajustement.*

d) - *Quelle pourrait être la taille d'un enfant de poids 2,8 kg à la naissance ? Peut-on estimer la taille d'un enfant ayant pour poids 2 kg à la naissance ?*

Ces activités permettent déjà de comprendre :

- qu'une étude au préalable assez précise du nuage s'impose ;
- qu'un domaine de validité doit accompagner le modèle ;
- et que des précautions s'imposent pour extrapoler en dehors du nuage.

6 - Démarche pour l'analyse d'une série bivariée

Une démarche peut alors être proposée aux élèves :

- REPRÉSENTER GRAPHIQUEMENT le nuage N(X , Y).
- ANALYSER le nuage N(X , Y) à l'aide de trois critères :
  - Structure : Existe-t-il des groupes de points bien séparés ou des points nettement isolés dans la direction générale de l'allongement du nuage ?
  - Linéarité : Le nuage s'allonge-t-il autour d'une droite ?
  - Homogénéité de l'épaisseur : Le nuage a-t-il une épaisseur homogène ?

Si un (ou plusieurs critères) n'est (ne sont) pas rempli(s) :

- partager le nuage en plusieurs groupes pour avoir des groupes homogènes, puis reprendre …
- trouver une explication pour les points isolés et les traiter à part si possible, puis reprendre …

Si aucune de ces techniques n'aboutit à un nuage présentant une réponse satisfaisante aux trois critères, l'ajustement affine ne peut pas être utilisé pour mettre en évidence une relation entre les caractères X et Y.

Il n'existe peut-être pas de relation ou bien les outils dont on dispose ne permettent pas de la mettre en évidence.

- TROUVER les valeurs des coefficients a et b de l'équation de la droite d'ajustement.

- INTERPRÉTER concrètement les coefficients *a* et *b* avec les unités.

  Lorsque x augmente de une unité, y augmente de *a* unités. *b* est la valeur de y lorsque x vaut 0.[4]
- RÉALISER des prévisions :
  - par interpolation sur le domaine de validité (à l'intérieur du nuage),
  - par extrapolation (à l'extérieur du nuage), en précisant les hypothèses nécessaires à la validité des calculs.
- COMPARER éventuellement plusieurs modèles issus de groupes différents.

II - Liens entre les deux caractères d'une série bivariée

1 - Introduction du problème

Les statistiques présentées en statistique descriptive peuvent ne concerner qu'un seul caractère : chaque individu (ou unité statistique) donne lieu à la mesure ou au relevé d'une valeur.

La situation qui va être envisagée ici est celle où l'on observe *deux* caractères dans une population : chaque individu donne lieu à la mesure ou au relevé de deux valeurs.

Soient *X* et *Y* les deux caractères considérés sur une population de *n* individus ; pour chaque individu, leur observation donne lieu à un *couple de valeurs* (x_i , y_i). Les données se présentent donc sous la forme d'une série statistique double présentée le plus souvent en tableau à deux colonnes ou à deux lignes.

Dans la suite, de même que dans notre exemple introductif, on note *x* la série statistique des $(x_i)_{1\leq i\leq n}$, *y* la série des $(y_i)_{1\leq i\leq n}$ et (x , y) la série bivariée $(x_i , y_i)_{1\leq i\leq n}$.

Nous ne considérerons ici que des *caractères quantitatifs*.

Ainsi peut-on envisager pour *X* et *Y* séparément, les résumés statistiques de position et de dispersion, notamment les moyennes des observations notées $\overline{x}$ et $\overline{y}$, et les écarts-types notés $s(x)$ et $s(y)$. Mais ces paramètres ne rendent pas compte d'une possible relation entre *X* et *Y* dans la population étudiée.

Les deux caractères ne jouant pas en général des rôles symétriques dans l'étude, on considère parfois que le caractère *X* est le *caractère explicatif* (ou indépendant) et le caractère *Y,* le *caractère à expliquer* (ou dépendant).

Cette notion d'*explication* n'est pas à confondre avec la notion de *causalité* qui nécessite, elle, une connaissance de la nature des mécanismes entrant en jeu (voir VI). Elle relève d'un choix intentionnel qui hiérarchise les caractères *X* et *Y* en fonction de la question posée sur le degré de leur interdépendance.

---

[4] Dans un premier temps, pour des raisons pédagogiques, il m'a semblé souhaitable de ne pas parler de *validation du modèle par les résidus*, ni de *mesure de la qualité d'ajustement avec* r^2, afin de mieux se concentrer sur l'analyse directe du nuage et la validation du modèle qui en découle.

2 - Représentation graphique : le nuage de points (diagramme de dispersion)

Pour visualiser au mieux une série statistique double, une représentation graphique dans le plan rapporté à un repère est une aide précieuse. Ainsi, comme nous l'avons vu dans l'exemple introductif, la première étape dans l'étude de la corrélation est la représentation graphique des données sous forme d'un *nuage de points* : à chaque individu i, on fait correspondre dans un repère cartésien, un point de coordonnées (x_i , y_i). Bien entendu les axes seront gradués selon des échelles qui tiennent compte des valeurs observées.

Si les deux caractères jouent des rôles différents, on mettra impérativement sur l'axe des abscisses le caractère explicatif et sur l'axe des ordonnées le caractère à expliquer. Par exemple, dans le cas traité en introduction, on a considéré que le caractère explicatif était le *revenu* et le caractère à expliquer était le *budget voiture* (en fonction du salaire du ménage). Dans le cas de caractères jouant des rôles symétriques, on choisira de placer les caractères X et Y arbitrairement en abscisses ou en ordonnées, mais les conclusions devront en tenir compte.

Remarque : Le point G de coordonnées $(\bar{x} , \bar{y})$ est appelé *point moyen du nuage*.

3 - Paramètre d'une série statistique double : covariance entre *X* et *Y*

La recherche d'une relation entre les caractères X et Y fait intervenir un indicateur statistique basé sur les écarts simultanés des valeurs observées x_i et y_i sur chaque individu à leurs moyennes. Cet indicateur est la covariance de la série statistique double, que l'on note $\text{cov}(x , y)$.

Définition

On appelle covariance d'une série double $(x , y) = (x_i , y_i)_{1 \leq i \leq n}$ la moyenne des produits des écarts des valeurs observées à leurs moyennes :

$$\text{cov}(x , y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Nous verrons que cet indicateur ressort dans plusieurs formules par la suite.

On peut remarquer que la covariance est augmentée lorsque les deux écarts relatifs à un même individu sont de même signe ; elle diminue lorsque les écarts sont de signes opposés. La covariance est donc, en général, positive s'il y a beaucoup d'écarts de même signe et négative s'il y a beaucoup d'écarts de signes contraires.

Calcul pratique de la covariance

Lorsque les moyennes ne sont pas des nombres entiers, il est souvent préférable, pour simplifier les calculs, de transformer la formule de la covariance en développant le produit $(x_i - \bar{x})(y_i - \bar{y})$ et en remarquant que $\sum_{i=1}^{n} x_i = n\bar{x}$ par exemple.

Après simplification on obtient alors la formule généralisée de König-Huygens suivante, désignant par *xy* la série statistique des produits $x_i y_i$:

$$\mathrm{cov}(x\,,\,y) = \overline{xy} - \bar{x}\,\bar{y}.$$

On peut dire brièvement que la covariance est égale à la moyenne des produits moins le produit des moyennes**.**

III - Ajustement affine par la méthode des moindres carrés

1 - La notion d'ajustement et son but

On cherche à mettre en évidence une relation qui existerait entre les deux caractères *X* et *Y* mesurés par une série double ; on cherche une fonction dont la courbe représentative ajuste au mieux le nuage, c'est-à-dire qui passe *au plus près possible* des points du nuage. Si la forme du nuage le suggère, nous préciserons comment, la fonction cherchée sera de type affine ; c'est le modèle explicatif le plus simple. On parle alors d'*ajustement affine*.

Si l'équation réduite de la droite d'ajustement est $y = a\,x + b$, l'observation d'une valeur x_0 du caractère *X* permet de prévoir la valeur qui devrait être prise par *Y*, à savoir : $\hat{y_0} = a\,x_0 + b$. En notant $\hat{y_i} = a\,x_i + b$, on appelle *résidu* $e_i = y_i - \hat{y_i}$, l'écart entre la valeur réelle y_i et la valeur calculée $\hat{y_i}$ à partir de x_i à l'aide de l'équation de la droite d'ajustement. Si on trace la droite d'équation $y = a\,x + b$, les résidus e_i sont représentés sur le graphique par des segments de droites verticaux, *d'abscisses* x_i, situés au-dessus ou en dessous du point de coordonnées $(x_i\,,\,y_i)$, suivant le signe de e_i.

Dans le cas de l'ajustement affine, on propose deux méthodes classiques : la méthode des points moyens due à MAYER, utilisée dans l'exemple introductif, et la méthode des *moindres carrés*, que l'on va exposer maintenant.

2 - Droite des moindres carrés

Principe

Soit $(x_i\,,\,y_i)_{1\le i\le n}$ une série statistique double représentée par son nuage de points. Parmi les droites d'équation $y = a\,x + b$, on cherche quelle est *la meilleure* (selon un certain critère) pour représenter le nuage ; autrement dit, il s'agit de déterminer *a* et *b* de telle sorte que la droite approche *au mieux* les différents points du nuage.

Cette droite d'ajustement peut être déterminée par la méthode des moindres carrés, c'est-à-dire de manière à rendre minimale la somme des carrés des résidus. La somme des carrés des écarts à minimiser, notée $SCE_{RÉSIDUELLE}$ en abrégé, est :

$$SCE_{RÉSIDUELLE} = \sum_{i=1}^{n} \left(y_i - \hat{y_i}\right)^2 = \sum_{i=1}^{n} \left[y_i - (a\,x_i + b)\right]^2$$

La résolution de ce problème permet de trouver les coefficients a et b donnant la meilleure droite d'ajustement appelée *droite d'ajustement affine de y en x associée à la série double* $(x_i , y_i)_{1\leq i\leq n}$, ou plus simplement *droite d'ajustement affine de y en x.*

On peut énoncer le résultat suivant :

La droite d'équation $y = a\,x + b$ *qui rend la* $SCE_{RÉSIDUELLE}$ *minimale est la droite qui passe par le point moyen* $G(\overline{x}, \overline{y})$ *et qui a pour coefficient directeur :* $a = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \dfrac{cov(x\,;\,y)}{Var(x)}$

Cette droite s'appelle la droite d'ajustement affine des moindres carrés de y en x, *et a donc pour équation :* $y = a\,(x - \overline{x}) + \overline{y}$.

3 - Droite d'ajustement affine des moindres carrés de x en y - Coefficient de corrélation linéaire

Il y a souvent une raison tenant à la nature des données pour étudier l'ajustement de y en x : dans de nombreux cas en effet, les deux caractères ne jouent pas le même rôle dans l'observation ; on cherche à *expliquer* les valeurs de Y par celles de X. Par exemple, si l'on relève l'âge et la taille d'enfants pour dresser des courbes de croissance, c'est l'âge qui s'impose comme caractère explicatif X.

Mais il en est pas toujours ainsi (exemple : taille à la naissance et taille adulte). De plus il est intéressant, comme nous allons le voir, de comparer l'ajustement de y en x déjà fait, avec l'ajustement de x en y (rôles inversés de X et Y).

La droite d'ajustement de x en y a pour équation $x = \alpha\,y + \beta$ *avec :*

$$\alpha = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = \frac{cov(x\,;\,y)}{Var(y)} \quad et \quad \beta = \overline{x} - \alpha\,\overline{y}.$$

D'où $x = \alpha\,(y - \overline{y}) + \overline{x}$. Cette droite passe aussi par le point moyen $G(\overline{x}, \overline{y})$.

Remarque : Les coefficients a et α sont de même signe, celui de cov(x , y), c'est-à-dire celui de $\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$.

Proposition

Si le nuage associé à une série statistique double est constitué de points alignés sur une droite d'équation $u\,x + v\,y + w = 0$, *avec* u *et* v *non nuls, alors cette droite est à la fois la droite d'ajustement de y en x et la droite d'ajustement de x en y.*

Dans le cas où le nuage n'est pas constitué de points alignés, les deux droites d'ajustement diffèrent. Mais l'appréciation de l'écart angulaire entre les deux droites demande de fixer les échelles en x et en y.

Le choix des unités peut perturber la lecture et l'interprétation graphique !

Souvent aucun choix d'échelle ne s'impose à priori : c'est le cas chaque fois que X et Y se réfèrent à des grandeurs de nature différente.

C'est pourquoi, on peut convenir d'un *choix d'unités normalisées*, revenant à un changement de variables décrit ci-après. En repère orthonormé, après ce changement, l'écart angulaire entre les deux droites d'ajustement aura un sens intrinsèque.

Nuage en variables centrées et réduites

On suppose que $s(x)$ et $s(y)$ sont non nuls, ce qui revient à dire que les x_i, tout comme les y_i ne prennent pas tous la même valeur.

Considérons alors la série statistique $(x', y') = (x_i', y_i')_{1\leq i\leq n}$ définie par :

$$x_i' = \frac{(x_i - \bar{x})}{s(x)} \text{ et } y_i' = \frac{(y_i - \bar{y})}{s(y)}$$

On dit que x_i' et y_i' ont été obtenus à partir de x_i et y_i par centrage et réduction.

On obtient alors pour le nouveau nuage, les coefficients a' et b' de la droite d'ajustement de y' en x', les séries images :

$$b' = 0 \text{ et } a' = \frac{1}{n}\sum_{i=1}^{n} x_i' y_i' = \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i - \bar{x})}{s(x)} \frac{(y_i - \bar{y})}{s(y)}.$$

On en déduit le théorème suivant (même étude pour l'ajustement de x' en y') :

Théorème :

Soient $(x_i , y_i)_{1\leq i\leq n}$ une série statistique double et $(x_i' , y_i')_{1\leq i\leq n}$ la série transformée par centrage et réduction des deux séries x et y. Cette série transformée admet pour droite d'ajustement de y' en x', la droite d'équation :

$$y = \frac{1}{n}\left(\sum_{j=1}^{n} x_i' y_i'\right) x = \frac{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}).(y_i - \bar{y})}{s(x).s(y)} x$$

et pour droite d'ajustement de x' en y', la droite d'équation :

$$x = \frac{1}{n}\left(\sum_{j=1}^{n} x_i' y_i'\right) y = \frac{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}).(y_i - \bar{y})}{s(x).s(y)} y.$$

Dans le plan, les deux droites d'ajustement du théorème précédent ont des coefficients directeurs inverses l'un de l'autre. En effet, si l'équation de la droite

d'ajustement de y' en x' est de la forme $y = r\,x$ et si $r \neq 0$, celle de x' en y' est de la forme $y = \frac{1}{r}x$.

Définition

Le coefficient $r = \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i - \bar{x}).(y_i - \bar{y})}{s(x).s(y)}$, *coefficient directeur de la droite d'ajustement de y' en x' obtenue après centrage et réduction, est appelé le* coefficient de corrélation linéaire[5] *dans la série statistique bivariée (x , y).*

Cette définition montre que r ne change pas si l'on change d'unités pour exprimer X et Y.

Interprétation graphique : Droites d'ajustement pour la série statistique *après centrage et réduction.*

Dans un repère cartésien orthonormé, les deux droites d'ajustement sont symétriques par rapport à la première bissectrice.

Si $r = 0$, elles sont orthogonales, confondues avec les axes de coordonnées. C'est le cas de la non corrélation entre x et y dans la série bivariée.

Elles ne peuvent être confondues que si $r^2 = 1$, c'est le cas de l'alignement parfait des points du nuage N(X, Y). Les séries x' et y' sont alors proportionnelles, de rapport positif quand $r = 1$ (x_i et y_i varient dans le même sens) et négatif quand $r = -1$ (x_i et y_i varient en sens contraire).

Quand les points de coordonnées (x_i , y_i) sont alignés horizontalement (cas où $s(y) = 0$), ou verticalement (cas où $s(x) = 0$), le coefficient de corrélation r n'est pas défini.

On voit ainsi que le coefficient de corrélation linéaire est un paramètre qui permet d'apprécier la qualité de l'ajustement d'une droite au nuage.

Remarque : On vérifie qu'après centrage et réduction l'angle θ que font entre elles les deux droites d'ajustement vérifie : $\cos\theta = \frac{2r}{1+r^2}$.

Exemple :

L'équation de la droite d'ajustement de y en x, selon la méthode des moindres carrés, pour les données salaire - budget voiture*, est* $y = 6{,}3194\,x - 1\,794{,}3488$, *alors que la droite de Mayer précédemment définie avait pour équation* $y = 6{,}49\,x - 2\,184{,}87$.

[5] La définition du coefficient de corrélation fait jouer des rôles symétriques aux observations des caractères X et Y. Sa valeur ne fait que mesurer le degré de pertinence d'une relation affine modélisant leur interdépendance et ne signifie en rien une relation de causalité.

– On pourra traiter cet exemple avec une calculatrice ou avec un tableur sur ordinateur (résultats donnés en annexe 1).

– On trouvera d'autres exemples dans l'annexe 2[6].

IV - Mesure de la qualité de l'ajustement - validation du modèle affine

Il est clair que l'on peut toujours proposer une droite d'ajustement quelle que soit la forme du nuage de points, même si l'ajustement affine ne se justifie pas.

Donc il est important d'insister sur *l'analyse graphique* du nuage selon les trois critères énoncés précédemment.

Il est aussi intéressant d'accompagner ce travail d'un indicateur – ou d'un coefficient – qui mesure la *qualité de l'ajustement*.

Le calcul de paramètres nouveaux permet alors d'évaluer une certaine qualité de cet ajustement (la part du facteur X dans la variabilité de Y).

1 - Équation d'analyse de variance

a) - Quelques définitions

$y_i - \bar{y}$ est l'*écart total* de y_i à la moyenne.

$\hat{y}_i - \bar{y}$ est *l'écart* à la moyenne de y_i *expliqué* par l'ajustement affine (dû au modèle).

On a vu que $y_i - \hat{y}_i$ est l'*écart résiduel* de y_i ou *résidu.*

b) - D'une égalité évidente à une égalité fondamentale

$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$: écart total = écart expliqué + écart résiduel

Dans le cas du modèle affine par la méthode des moindres carrés, on démontre la relation suivante dite *équation d'analyse de variance* :

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Remarque : Cette égalité algébrique, qui pourra être vérifiée avec l'ordinateur lors du traitement informatique des problèmes, est une illustration *propre au modèle affine par la méthode des moindres carrés* de l'égalité plus générale suivante :

$$(1) \quad SCE_{TOTALE} = SCE_{MODÈLE} + SCE_{RÉSIDUELLE}$$

[6] On trouvera les fichiers tableurs téléchargeables sur le site de la Commission Inter IREM *Statistique et Probabilités*. On pourra consulter, pour d'autres exemples, l'article d'Hubert RAYMONDAUD : *Description des séries bivariées quantitatives* (volume 2).

2 - Coefficient de détermination

a) - Définition

Si $SCE_{TOTALE} \neq 0$, c'est-à-dire si la série statistique $(y_i)_{1\leq i\leq n}$ n'est pas constante, on peut diviser les deux membres de l'égalité (1) par SCE_{TOTALE}, on obtient :

$$1 = \frac{SCE_{MODÈLE}}{SCE_{TOTALE}} + \frac{SCE_{RÉSIDUELLE}}{SCE_{TOTALE}}.$$

Par définition, le coefficient de détermination est le rapport $\delta = \frac{SCE_{MODÈLE}}{SCE_{TOTALE}}$, dans le modèle d'ajustement affine par la méthode des moindres carrés, on a donc :

$$\delta = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}.$$

b) - Propriété et interprétation du coefficient de détermination

On a donc $0 \leq \delta \leq 1$.

Le coefficient de détermination est la proportion de la variabilité totale de Y qui est *expliquée* par l'ajustement affine de y en x. On l'exprime souvent en pourcentage.

Si on obtient un coefficient de détermination $\delta = 0{,}9$, cela signifie que 90 % des variations de Y sont expliquées par l'ajustement affine.

On remarque que cette interprétation donne tout son sens au modèle.

Dans notre exemple introductif, le coefficient de détermination entre les caractères *salaire* et *budget voiture* vaut 0,8968.

Le coefficient de détermination, défini dans le cas du modèle affine par la méthode des moindres carrés, est égal à r^2. En effet,

$$\frac{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2} = \frac{\sum_{i=1}^{n} \left(a\,x_i + b - a\,\bar{x} - b\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2} = a^2 \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}.$$

et, vu la valeur de a obtenue en III 2),

$$\frac{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2} = \left[\frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}\right]^2 \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2} = r^2.$$

c) - Appréciation de la pertinence d'un modèle affine

Lorsque les moyens de représentation graphique étaient peu développés – l'apparition des ordinateurs possédant des écrans et des périphériques graphiques a modifié l'approche de ce problème – le coefficient de corrélation était calculé au préalable pour donner une idée du degré d'alignement des points (quand r^2 tend vers 1, l'angle θ des deux droites d'ajustement tend vers 0 ou π d'après la formule de la fin du paragraphe III), et donc d'apprécier la pertinence du modèle affine, sans être obligé de tracer réellement le nuage de points. Cependant, cette démarche n'était pas sans risques, car un coefficient de corrélation même très proche de 1 ne garantit pas que le modèle affine soit le meilleur que l'on puisse considérer, comme le montre l'exemple de la relation poids-taille à la naissance, traité par l'analyse des résidus ci-dessous.

Aujourd'hui l'usage des logiciels statistiques et la possibilité de *voir* rapidement le nuage, permettent de procéder dans l'ordre inverse : l'œil permet de juger efficacement la pertinence de l'ajustement affine ; le calcul des indicateurs mesure la qualité de cet ajustement.

Il faut donc toujours se rappeler que l'œil est capable d'appréhender simplement et rapidement la complexité d'un nuage de points. Il voit en particulier si la relation est adaptée pour tous les points ou s'il existe des individus pour lesquels la relation n'est pas justifiée. Dans ce dernier cas, il est judicieux de commencer par *traiter* ces points avant de faire l'ajustement affine. L'œil voit aussi si la relation entre X et Y est proche d'une relation affine.

Application

On sélectionne 12 personnes inscrites à un stage de formation. Avant le début de la formation, les stagiaires subissent une épreuve A notée de 0 à 20 ; à l'issue du stage, une épreuve B identique à la première est notée aussi de 0 à 20. Les résultats sont rassemblés dans le tableau suivant :

Stagiaires	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Épreuve A	*3*	*4*	*6*	*7*	*9*	*10*	*9*	*11*	*12*	*13*	*15*	*4*
Épreuve B	*8*	*9*	*10*	*13*	*15*	*14*	*13*	*16*	*13*	*19*	*6*	*19*

On note X (respectivement Y) le caractère « note à l'épreuve A » (respectivement « note à l'épreuve B »).

a) - *Dessiner le nuage des points représentant les 12 couples de notes.*

b) - *Analyser le nuage selon les trois critères.*
Des points particuliers peuvent être soit gardés pour les calculs, soit retirés ; il faut alors justifier le choix qui est fait !

c) - *Déterminer alors l'équation de la droite d'ajustement affine.*

d) - *Interpréter concrètement, si cela a du sens, les coefficient a et b.*

Le coefficient de détermination ne permet pas de juger de la pertinence d'un modèle ; il permet seulement de mesurer la qualité de l'ajustement affine. Mais un autre ajustement pourrait se révéler plus performant. Donc l'étude de l'ajustement affine peut se prolonger par une recherche d'un meilleur modèle que le modèle affine, notamment par l'examen des résidus.

3 - Analyse des résidus, validation du modèle

Même si l'importance des résidus d'un modèle est limitée, il est toujours instructif de procéder à leur analyse afin de vérifier :

- s'ils ne révèlent pas une mauvaise spécification du modèle utilisé ;
- s'ils ne mettent pas en évidence l'existence d'autres variables explicatives que celle qui a été retenue.

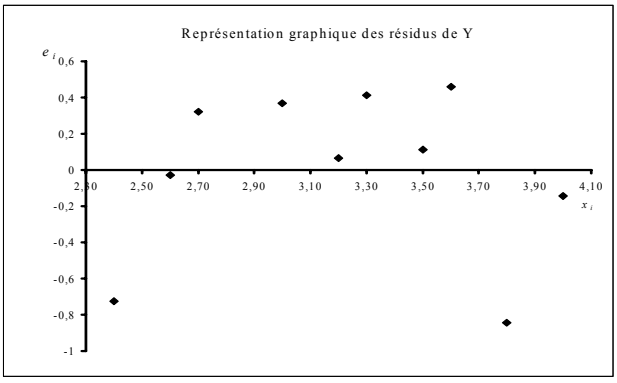
Concernant les résidus, dans le cas du modèle affine, leur somme est nulle. Dans une représentation graphique, ils se répartissent de façon aléatoire entre résidus positifs et négatifs, sans qu'aucune tendance n'apparaisse…

Ces critères permettent de valider le modèle choisi ; dans le cas contraire, cela met en évidence que le modèle n'est pas bien adapté.

Ces résultats ne s'appliquent que dans le cas du modèle affine.

Remarque : On a déjà signalé plus haut l'influence que pouvait avoir le choix des unités pour le repère dans lequel on construit le nuage de point ; et notamment fausser l'analyse du nuage selon les trois critères proposés en masquant une courbure par exemple.

Le diagramme des résidus peut permettre alors de révéler cette erreur d'interprétation graphique en montrant une répartition tendancielle des résidus. C'est particulièrement le cas dans l'exemple de la relation poids-taille à la naissance donné en activité 2 d'approche de l'ajustement affine.



Pour la corrélation entre poids et taille à la naissance, on obtient : $r^2 \approx 0{,}9828$.

La représentation graphique des résidus montre une certaine tendance : les résidus sont positifs *au centre* ; l'ajustement affine n'est manifestement pas le modèle le mieux adapté, malgré un coefficient de détermination proche de 1.

V - Que faire lorsque le nuage de points paraît s'étirer autour d'une courbe qui n'est pas une droite ?

Il arrive souvent qu'une relation nette apparaisse entre deux caractères sans que cette relation soit affine ; et il peut être utile de procéder à l'*ajustement* d'une courbe au nuage des points observés. La démarche générale adoptée est la suivante.

1 - Choix de l'équation de la courbe d'ajustement (modélisation)

Il est basé soit sur des considérations théoriques, soit sur des considération empiriques. D'une manière générale, l'utilisation de bases théoriques est préférable car on obtient à la fois une équation qui s'adapte bien à la réalité, mais aussi des paramètres intervenant dans cette équation qui possèdent une signification précise (voir 2 b) remarque ci-après !).

Il faut noter cependant que ces considérations théoriques conduisent la plupart du temps à une formalisation mathématique sous forme d'un système d'équations différentielles.

Dans la pratique, le type d'équation de la courbe peut donc être *choisi dans un catalogue* selon des connaissances théoriques a priori (exemple : fonction exponentielle pour un phénomène de croissance de population ou d'organismes assez jeunes), ou bien l'équation exacte de la courbe peut se trouver par résolution du système d'équations différentielles (voir 2 c).

Quand on ne dispose d'aucune base théorique satisfaisante, on peut rechercher le type d'équation de courbe à ajuster en fonction des données observées elles-mêmes, c'est-à-dire essayer de façon empirique de coller au mieux au nuage de départ. Par exemple :

- en transformant les données de manière à se ramener à un ajustement affine (en utilisant par exemple l'échelle proposée par TUKEY dans *Exploratory Data Analysis*, où le degré de la fonction de transformation dépend de la concavité de la courbe autour de laquelle s'allonge le nuage) ;
- en ajustant une équation polynomiale de degré k suffisamment élevé ;
- en ajustant des fonctions différentes en vue d'identifier celle que l'on pourrait qualifier de "meilleure".

2 - Détermination des paramètres qui interviennent dans cette courbe (ajustement proprement dit)

Les trois principales méthodes que l'on peut utiliser sont[7] :

- La méthode du changement de variable(s) permettant de se ramener à un *ajustement affine* **;**
- Les méthodes d'optimisation non linéaire **;**
- La méthode de formalisation conduisant à un système *d'équations différentielles.*

a) - Ajustement qui, par un changement de variable, se ramène à un ajustement affine

Cette méthode consiste à faire un changement de variable pour X ou pour Y ou pour les deux, permettant d'obtenir, par transformation, un nuage pour lequel le modèle affine est pertinent.

Le choix de la transformation peut se faire empiriquement (voir ci-dessus avec l'échelle de TUKEY), soit par implication, suivant l'équation de la courbe d'ajustement qui a été préalablement choisie (par exemple, si l'on choisit comme équation pour la courbe d'ajustement $y = c\,\mathrm{e}^{a\,x}$, on fera le changement de variable $Z = \ln Y$).

On procède alors à l'estimation des paramètres de la fonction affine selon le critère des moindres carrés, à partir des variables images.

Par exemple, dans le cas où le changement de variable porte sur Y, on ajuste le modèle $f(Y)_{estimé} = a\,X + b$.

Le critère des moindres carrés permet d'obtenir la plus petite $SCE_{RÉSIDUELLE}$ relativement à la variable $f(Y)$.

Que se passe-t-il lorsque l'on revient à la variable Y ?

Pour revenir à la variable Y, on utilise la fonction réciproque de f. Le modèle ajusté est alors $Y_{estimé} = f^{-1}([f(Y)]_{estimé})$. Il est important de savoir qu'alors, la $SCE_{RÉSIDUELLE}$ obtenue pour $Y_{estimé}$ n'est plus la plus petite possible. L'exemple du problème de maintenance qui suit ce paragraphe le montre.

En effet, les équations normales ont porté sur $f(Y)$ et non sur Y, c'est-à-dire que la $SCE_{RÉSIDUELLE}$ par rapport à la droite d'ajustement est minimisée en fonction des valeurs transformées et non en fonction des valeurs initiales.

Remarques :

i - Une situation où la méthode des moindres carrés peut être appliquée de manière directe, sans changement de variables, et conduit à un système d'équations normales (linéaires en les $k + 1$ paramètres), est celle de l'*ajustement par une*

[7] Les ouvrages de E. JOLIVET (1983), et de J. D. LEBRETON et C. MILLIER (1982) donnent plusieurs exemples simples de l'application de ces trois méthodes en biologie.

équation polynomiale de degré *k*. On peut parler alors d'*ajustement non affine* par la méthode des *moindres carrés linéaires.*

ii - Le changement de variable était rendu nécessaire par l'absence d'algorithmes d'ajustement non affine selon les moindres carrés et/ou par l'accès difficile aux programmes permettant de les mettre en œuvre. Actuellement, des algorithmes professionnels (dont l'optimalité vis à vis de la $SCE_{RÉSIDUELLE}$ est démontrée mathématiquement) sont disponibles facilement sur Internet (par exemple R, voir ci-après).

C'est pour cela que la méthode du changement de variable n'est quasiment plus utilisée dans la pratique professionnelle du traitement statistique et a été remplacée par l'utilisation d'algorithmes d'optimisation dite non linéaire.

b) - Méthodes d'optimisation non linéaire

Elles consistent à utiliser des algorithmes d'ajustement non affine, le plus couramment basés sur le critère des moindres carrés, appelé alors moindres carrés non linéaires (ici non seulement l'équation de la courbe d'ajustement est non affine, mais les équations normales sont alors non linéaires en les paramètres).

Par exemple l'algorithme de MARQUARDT (1963)[8] fut largement utilisé. Un autre plus moderne, l'algorithme « nls() » de R, est facilement accessible.[9]

On peut donc ajuster directement, par exemple, une fonction $Y_{\text{estimé}} = e^{(ax+b)}$. L'algorithme permet d'obtenir la plus petite $SCE_{RÉSIDUELLE}$ relativement à ce modèle. Cette $SCE_{RÉSIDUELLE}$ est plus petite que celle obtenue à partir de $f^{-1}([f(Y)]_{estimé}$.[10]

Remarques :

i - Pour un ajustement affine ou polynomial d'ordre $k > 1$, ces algorithmes sont strictement équivalents à la méthode des moindres carrés linéaires.

ii - Si pour une fonction donnée, la $SCE_{RÉSIDUELLE}$ minimale est toujours un critère pertinent pour calculer les valeurs des paramètres, elle ne l'est plus systématiquement pour le choix d'un type de fonction parmi plusieurs autres. Il est préférable de choisir un modèle dont les paramètres ont une interprétation

---

[8] Le programme (HAUSS59) de cet algorithme, utilisé par diverses unités du département de biométrie de l'I.N.R.A. (TOMASSONE et ROUX 1973), est détaillé en annexe p. 135 de l'ouvrage de E. JOLIVET [1983].

[9] R est un langage professionnel de programmation pour l'usage des statistiques. Il est disponible gratuitement sur Internet, en versions Windows, Linux, Mac, sur le site www.r-project.org. Ses références sont : R : *A Programming Environment for Data Analysis and Graphics, 1999–2004. R Development Core Team from the R-project*.
Les références relatives à l'algorithme nls() (qui est une fonction de R, au sens informatique du terme) sont : BATES D.M., WATTS D.G. [1988] et BATES D. M., CHAMBERS J. M. [1992].

[10] Attention : les méthodes utilisées par les menus *ajustements exponentiel, logarithmique, puissance...* des calculatrices, ne sont, actuellement, que des ajustements affines après changement de variables. Les algorithmes d'optimisation non linéaire n'y sont pas encore implantés.

concrète quant au phénomène étudié, plutôt qu'un autre donnant un $SCE_{RÉSIDUELLE}$ plus petite, mais dont les paramètres n'ont pas d'interprétation concrète simple.

Par exemple, dans l'étude de la croissance pondérale d'une population bactérienne (voir l'activité pluridisciplinaire de l'annexe 2), il est préférable de choisir un modèle exponentiel, qui est issu de la connaissance des dynamiques de croissance de ces populations, et qui fournit des paramètres interprétables concrètement, plutôt qu'un modèle polynomial, qui fournit une plus petite $SCE_{RÉSIDUELLE}$ mais dont les paramètres n'ont pas d'interprétation concrète.[11]

Les nuages de points rencontrés dans les exemples de cet article, peuvent tous être ajustés par des polynômes de degré 3 ou plus, avec des $SCE_{RÉSIDUELLE}$ plus petites que celles obtenues avec des modèles plus pertinents. C'est l'illustration d'une propriété générale des modèles polynomiaux. De façon imagée : il suffit de monter suffisamment haut en degré pour ajuster une fonction polynôme de degré k à n'importe quelle forme de courbe, et obtenir une très bonne $SCE_{RÉSIDUELLE}$.

Le tableau qui suit présente les valeurs des $SCE_{RÉSIDUELLE}$ pour différents modèles d'ajustement de la série double donnée en exemple de l'activité pluridisciplinaire (annexe 2). Les ajustements polynomiaux semblent performants, mais ne donnent pas lieu à une interprétation claire immédiate.

Modèle	**Paramètres et estimations diverses obtenus avec R**	$SCE_{RÉSIDUELLE}$
$\{\ln(\text{Masse})\}_{\text{estimé}} = at + b$	$a \approx 0{,}1357858$ est une estimation de la vitesse spécifique de croissance $b \approx 0{,}6456694$ est une estimation de la masse initiale, à $t = 0$	4,78205
$\text{Masse}_{\text{estimée}} = e^{(at+b)}$	$a \approx 0{,}1357858$ est une estimation de la vitesse spécifique de croissance $b \approx 0{,}6456694$ est une estimation de la masse initiale, à $t = 0$	0,9009527
$\text{Masse}_{\text{estimée}} = at^3 + bt^2 + ct + d$	$d \approx 0{,}420183007$ est une estimation de la masse initiale, à $t = 0$ $c \approx 0{,}168389565$ est une estimation de la vitesse de croissance à $t = 0$ $b \approx 0{,}012767745$ Quelle signification ? $a \approx 0{,}001199121$ Quelle signification ?	0,04104726
$\text{Masse}_{\text{estimée}} = at^4 + bt^3 + ct^2 + dt + e$	$a \approx 0{,}000009477308$??? $b \approx 0{,}0006683919$: ??? $c \approx 0{,}003414995$??? $d \approx 0{,}1145585$??? $e \approx 0{,}4722248$ est une estimation de la masse initiale, à $t = 0$	0,02356880

[11] Les instructions et les résultats obtenus avec R, figurent dans le document *R_Ajustements.doc* téléchargeable sur le site de la Commission Inter-IREM *Statistique et Probabilités*.

c) - Formalisation conduisant à des systèmes d'équations différentielles

Toutes les méthodes vues précédemment procèdent cependant toujours du *principe du catalogue* : d'après la forme du nuage, ou selon des connaissances théoriques sur le phénomène étudié, on choisit un type de fonction dont les courbes sont susceptibles de bien résumer/modéliser la relation illustrée par le nuage ; cette appréciation ou ce choix, même basé sur des considérations théoriques, entraîne des imprécisions qui peuvent être corrigées par une méthode de modélisation basée sur une formalisation mathématique détaillée (et souvent complexe) des mécanismes en jeu.

Cette troisième méthode se base alors sur la connaissance, lorsqu'elle est possible, et la modélisation des mécanismes élémentaires intervenant dans le phénomène étudié. Elle ressemble fort en cela, par exemple, à la façon dont un physicien assimile une trappe immergée à un rectangle et le découpe en fines bandes horizontales qu'il suppose soumises à une pression constante, de façon à pouvoir ensuite intégrer sur la hauteur de la trappe pour trouver la force à laquelle elle est soumise. Cette méthode aboutit à la résolution de systèmes d'équations différentielles, en général compliqués, et que l'on résout souvent par des méthodes numériques.

3 - Contrôle a posteriori de l'adéquation du modèle choisi (validation)

- Ce contrôle peut être réalisé, comme pour l'ajustement affine direct, par l'examen des résidus et par le calcul de la $SCE_{RÉSIDUELLE}$.
- Si l'on utilise une transformation, a et b se calculent à partir de la série double *image*, leur interprétation ne fournit plus d'indications pratiques simples, r^2 *ne représente plus la qualité de l'ajustement au nuage* (X, Y) *mais seulement celle au nuage image.* Dans le cas de l'ajustement exponentiel d'une série $(x_i, y_i)_{1\leq i\leq n}$, l'ajustement affine s'applique à la série transformée $(x_i, z_i)_{1\leq i\leq n}$ avec $z_i = \ln y_i$ et bien entendu les résultats concernant les résidus s'appliquent à cette série mais pas à la série $(x_i, y_i)_{1\leq i\leq n}$.

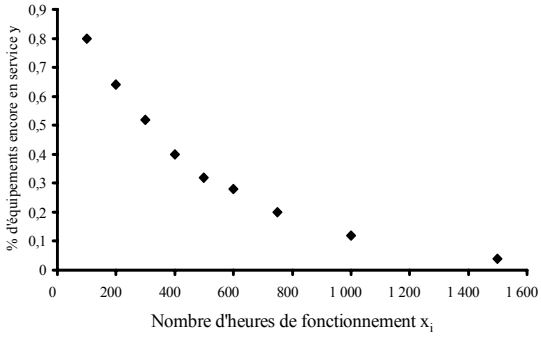
Pour comparer plusieurs modèles, on compare leurs $SCE_{RÉSIDUELLE}$ (relatives aux données initiales pour être dans les mêmes unités) et/ou leurs valeurs respectives du ratio précédent (utilisation de l'outil informatique) ; puis on peut analyser la représentation graphique des résidus.

Activité : Un problème de maintenance

On a étudié la durée de vie d'un certain nombre d'équipements mécaniques identiques. Dans le tableau suivant, x_i *représente le nombre d'heures de fonctionnement et* y_i *le pourcentage d'équipements encore en service au bout de* x_i*.*

x_i	*100*	*200*	*300*	*400*	*500*	*600*	*750*	*1 000*	*1 500*
y_i	*0,8*	*0,64*	*0,52*	*0,4*	*0,32*	*0,28*	*0,2*	*0,12*	*0,04*

1) - *On a représenté ici le nuage de points $N(x_i , y_i)$.*



Indiquer un critère d'analyse du nuage qui ne rend pas pertinent un ajustement affine direct.

2) - *Des raisons théoriques poussent, dans ce cas, à faire le changement de variable $Z = ln(Y)$. En vous aidant de la fiche-étude type,*

a) - *Représenter le nouveau nuage des points (x_i , z_i) dans le plan muni d'un repère orthogonal. Échelle : 2 cm représentent 200 unités en abscisses et 0,5 unité en ordonnées !*

b) - *Justifier maintenant que l'on peut envisager un ajustement affine de ce nuage.*

c) - *Trouver alors, en utilisant la calculatrice, une équation de la droite d'ajustement affine de Z en X, par la méthode des moindres carrés. Valeurs arrondies des coefficients à 10^{-4} près !*

d) - *Évaluer la qualité de l'ajustement affine en donnant, et en commentant, le coefficient de détermination. Valeur arrondie à 10^{-4} près !*

e) - *Calculer les résidus dus au modèle $\hat{Z}$ pour chaque valeur x_i de X, et en faire une représentation graphique. Échelle : 2 cm représentent 200 unités en abscisses et 0,01 unité en ordonnées ! Comment exploiter cette représentation ?*

3) - *Donner alors une équation de la courbe d'ajustement exponentiel et représenter graphiquement cette courbe sur le graphique d'origine.*

4) - *Estimations / Prévisions*

a) - *Déterminer une estimation du pourcentage d'équipements encore en service au bout de 900 heures de fonctionnement.*

b) - *Déterminer une estimation de la valeur x_0 pour laquelle 36,78 % des équipements sont encore en service.*

Nota : En utilisant la procédure *nls* de R, on trouve pour l'exemple précédent (problème de maintenance), les résultats suivants :

Modèle	**Paramètres**	$SCE_{RÉSIDUELLE}$ **des** Y
Ajustement affine après changement de variable : $\hat{Z} = a_z X + b_z$ $\hat{Y} = e^{\hat{Z}} = e^{a_z X + b_z}$	$a_z = -0{,}00212$ $b_z = -0{,}02951$	0,00087
Ajustement non affine $\hat{Y} = e^{aX+b}$	$a = -0{,}00217$ $b = -0{,}01097$	0,00067
Ajustement polynomial de degré 3		0,00064

VI - Corrélation et causalité

Le but des chercheurs en quête de corrélation est le plus souvent une recherche de causalité. On voudrait, en cas de corrélation entre caractères statistiques, transformer le caractère explicatif en caractère *causal* ; dans notre exemple, il serait tentant de dire que la relation trouvée entre le salaire mensuel et le budget relatif à l'achat d'une voiture est une *relation de cause à effet*.

Seule la connaissance de la réalité des mécanismes en action permet de transformer la corrélation en causalité et en général seul le praticien peut franchir ce pas. La corrélation ne dépend que d'un calcul alors que la causalité, bien que s'appuyant sur l'existence d'une telle liaison entre les caractères, est du domaine de la sémantique. La corrélation peut trouver son explication ailleurs que dans la causalité : une coïncidence, une antériorité, une cause commune, etc[12]. D'où la nécessité d'une bonne connaissance du problème et des données pour aller au-delà d'une simple constatation statistique.

VII - Épilogue

Le lecteur aura remarqué que le mot *régression* a été évité… Il s'agit en fait d'un acte volontaire pour (comme le faisait remarquer le titre de cet article) être bien dans les clous de nos programmes. En effet, il s'agit bien de s'en tenir à :

- la représentation graphique du nuage de points et son analyse (servant de validation a priori dans le choix d'un modèle),
- la détermination de la courbe d'ajustement dite aussi de régression ou d'estimation,
- l'interprétation éventuelle des paramètres y intervenant,
- la mesure de la qualité de cet ajustement,

[12] On peut rappeler ici l'exemple donné par Albert JACQUARD : il existe une très forte corrélation entre le montant des impôts sur le revenu payé par les familles et la pratique des sports d'hiver. Suffirait-il d'augmenter les impôts pour permettre aux familles défavorisées d'aller davantage faire du ski ?

- l'étude des résidus pour un contrôle a posteriori,
- faire des prévisions (estimation déterministe ponctuelle).

Il s'agit ici de la première phase de ce que l'on appelle communément une *analyse de régression*, qui se poursuit généralement par de l'*inférence* (analyse de variance, tests,…), pour mesurer le degré de confiance des estimations faites à partir de la série, considérée alors comme un échantillon (X et Y étant alors des variables aléatoires).

Éléments de bibliographie

BATES D. M., WATTS D. G., (1988) : *Nonlinear regression analysis and its applications*, New-York, Wiley.

BATES D. M., CHAMBERS J. M., (1992) : *Nonlinear models. Chapter 10 of Statistical Models in S*, eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

BOURSIN J. L., (1991) : *Comprendre la statistique descriptive*, Paris, Armand Colin.

DAGNELIE P., (1998) : *Statistique, théorique et appliquée (Tomes 1 et 2)*, Paris et Bruxelles, De Boeck & Larcier.

DODGE Y., (1993) : *Statistique*, Dictionnaire encyclopédique, Paris, Dunod.

GRES (Groupe de Réflexion sur l'Enseignement de la Statistique) (1995-2000) : *Bulletins du GRES*, Ministère de l'Agriculture et de la Pêche, Toulouse-Auzeville, ENFA.

IREM de Strasbourg (1983) : *Mathématiques*, terminales C et E, analyse et statistiques, Strasbourg, Librairie Istra.

JOLIVET E., (1983) : *Introduction aux modèles mathématiques en biologie*, INRA, Actualités scientifiques et agronomiques. Paris, Masson.

LANNUZEL B., (1999) : *Probabilités et statistique, Cours et exercices corrigés*, Paris, Dunod.

LEBRETON J. D., MILLIER C., (sous la direction de) (1982) : Par E. JOLIVET, C. MILLIER, J. D. LEBRETON, A. PAVÉ, J. P. VILA, *Modèles dynamiques et déterministes en biologie*, Paris, Masson.

MARQUARDT D. W., (1963) : *An algorithm for least square estimation of non linear parameters*, S.I.A.M. J., **11**, 431-441.

TOMASSONE R., AUDRAIN S., LESQUOY-DE TURCKHEIM E., MILLIER C., (1992) : *La régression : nouveaux regards sur une ancienne méthode statistique*, Paris, Masson.

TOMASSONE R., ROUX C., (1973) : *Ajustements non-linéaires (HAUSS59)*. Note interne du Laboratoire de Biométrie du C.N.R.S.

TUKEY J. W., (1977) : *Exploratory data analysis*, Reading, Addison-Wesley.

Annexe 1 : Fiches-calcul du traitement informatique des deux exemples cités

Budget-voiture : analyse d'une série bivariée et ajustement affine

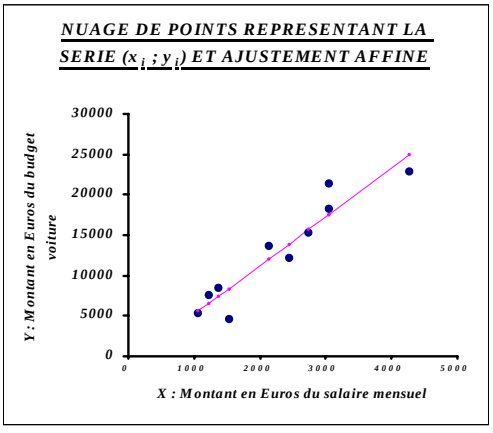
Identification des ménages :	Variable explicative X : Montant en euros du salaire mensuel	Variable expliquée Y : Montant en euros du budget voiture	Valeurs estimées de Y	Résidus de Y	Carrés des écarts résiduels	Carrés des écarts dus à la droite	Carrés des écarts totaux
7	1 067	5 350	5 612	– 262	68 869	53 986 794	57 912 100
1	1 220	7 600	6 534	1 066	1 136 114	41 292 015	28 729 600
5	1 372	8 400	7 450	950	902 930	30 362 592	20 793 600
2	1 524	4 550	8 365	– 3 815	14 557 537	21 110 036	70 728 100
9	2 134	13 700	12 040	1 660	2 755 192	846 174	547 600
6	2 439	12 200	13 877	– 1 677	2 813 897	841 746	577 600
4	2 744	15 250	15 715	– 465	216 050	7 588 988	5 244 100
3	3 049	18 300	17 552	748	559 270	21 087 900	28 515 600
8	3 049	21 350	17 552	3 798	14 423 616	21 087 900	70 392 100
10	4 269	22 900	24 902	– 2 002	4 006 139	142 600 239	98 803 600
Moyennes		12 960					
Sommes		129 600	129 600	0,000			
Sommes des carrés					41 439 614	340 804 386	382 244 000,000

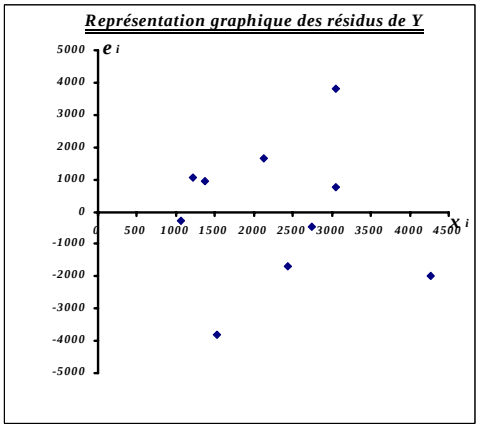
Prévision : On vérifie l'équation d'analyse de variance ! !
$SCE_{MODÈLE}/SCE_{TOTALE} = 0{,}8916$

Coefficients de la droite d'ajustement affine ; coefficient de corrélation ; coefficient de détermination

a	*b*	*r*	r^2	
6,0241	- 815,2642	0,9442	0,8916	89,16 % *des variations de Y sont expliquées par les variations de X*

L'équation réduite de la droite est : y = 6,0241 x – 815,2642 *pour* x *appartenant à l'intervalle* [1 067 ; 4 269]





Problème de maintenance : analyse d'une série bivariée, linéarisation et ajustement affine

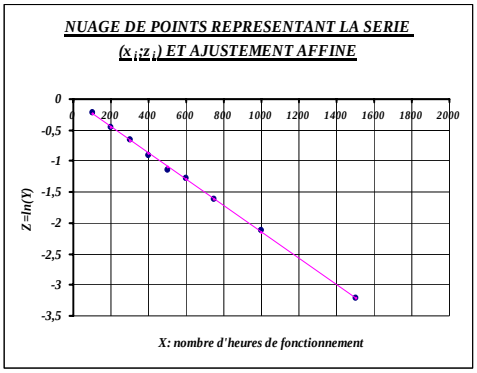
Identification des couples :	Variable explicative X : Nombre d'heures de fonctionnement	Variable expliquée Y : État du parc ; % en service	Z = ln(Y)	Valeurs estimées de ln(Y)	Valeurs estimées de Y	Résidus de Z	Carrés des écarts résiduels	Carrés des écarts dus au modèle	Carrés des écarts totaux
1	100	0,8	– 0,2231	– 0,2414	0,7855	0,018	0,000209088	0,17359825	0,18585679
2	200	0,64	– 0,4463	– 0,4533	0,6356	0,007	1,97361E-05	0,07111213	0,07350123
3	300	0,52	– 0,6539	– 0,6651	0,5142	0,011	3,35136E-05	0,02111848	0,02283456
4	400	0,4	– 0,9163	– 0,8770	0,4160	– 0,039	0,000257056	0,00222256	0,00096790
5	500	0,32	– 1,1394	– 1,0889	0,3366	– 0,051	0,000275562	0,00104256	0,00239012
6	600	0,28	– 1,2730	– 1,3007	0,2723	0,028	5,87793E-05	0,00932299	0,00790123
7	750	0,2	– 1,6094	– 1,6185	0,1982	0,009	3,27844E-06	0,02913833	0,02852345
8	1 000	0,12	– 2,1203	– 2,1482	0,1167	0,028	1,09333E-05	0,06360254	0,06194567
9	1 500	0,04	– 3,2189	– 3,2076	0,0405	– 0,011	2,07577E-07	0,10786842	0,10816790
Moyennes		0,37							
Sommes						– 0,018			
Sommes des carrés							0,001	0,479	0,492
Prévision :	900			– 1,9363	**0,1442**	On remarque au passage que l'équation d'analyse de variance n'est pas vérifiée !!			
	486	0,3678	– 1,0002						

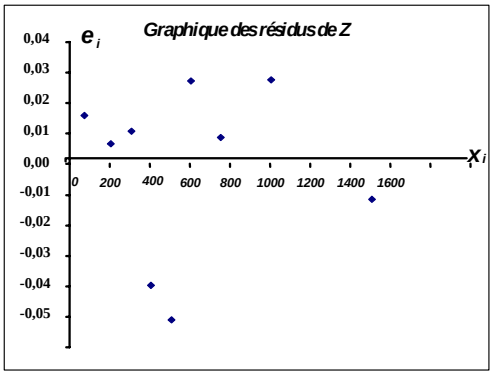
Coefficients de la droite d'ajustement affine ; coefficient de corrélation ; coefficient de détermination

a	*b*	*r*	r^2	
– 0,00212	– 0,02951	– 0,9995	0,9991	*99,9 % des variations de Z sont expliquées par les variations de X*

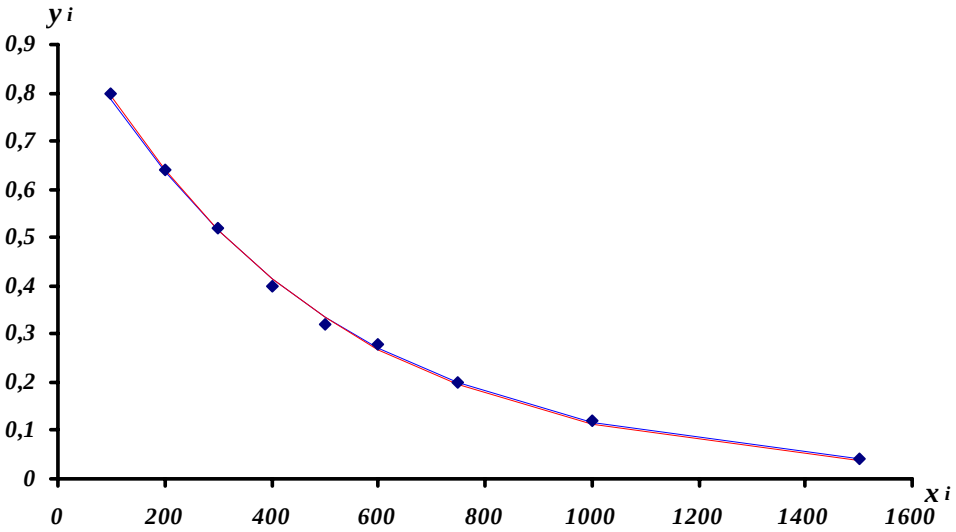
L'équation réduite de la droite est :

z = – 0,0021 x – 0,0295 *, pour* x *appartenant à l'intervalle* [100 ; 1 500]





Dans les documents suivants, figurent les deux ajustements exponentiels (indirect, obtenu par fonction réciproque après transformation, et direct, obtenu avec R).



Représentation du nuage (x_i , y_i) et ajustements exponentiels

Ajustement exponentiel indirect :
$y = \exp(-0{,}00212\,x - 0{,}02951)$
pour x appartenant à l'intervalle [100 ; 1 500]
$\frac{SCE_T - SCE_R}{SCE_T} = 0{,}9982$
$SCE_{RÉSIDUELLE}$ indirecte 0,0008

Ajustement exponentiel direct :
$y = \exp(-0{,}00217\,x - 0{,}01097)$,
pour x appartenant à l'intervalle [100 ; 1 500]
$\frac{SCE_T - SCE_R}{SCE_T} = 0{,}9986$
$SCE_{RÉSIDUELLE}$ directe 0,00067

Annexe 2 : Exemples d'exercices d'application, d'activités, de devoirs

Exercice 1.

[Le fichier-calcul disponible sur le site de la Commission Inter-IREM *Statistique et Probabilités* est *presatmoscorrigé.xls*]

Sur une même verticale, les hauteurs barométriques diminuent avec l'altitude, conformément au tableau :

Altitude *x* en km	0	1	2	4	6	10
Hauteur barométrique *y* en cm de mercure	76	67	59	46	35	20

1° a) - Tracer le nuage de points représentatif du tableau.

b) - Analyser le nuage quant aux 3 critères.

2°) - Si c'est envisageable, trouver une équation de la droite d'ajustement affine, par la méthode des moindres carrés ; la tracer sur le même graphique.

3°) - Donner l'interprétation concrète des coefficients a et b de l'équation.

4°) - Faire l'estimation de la pression atmosphérique au sommet du Ventoux (1 912 m).

Quelle pression atmosphérique peut-on prévoir à 15 km d'altitude ?

On précisera dans chacun des cas, les hypothèses nécessaires pour valider les calculs.

5°) - Estimer à partir de quelle altitude cette pression devient nulle (donc plus d'air !). *On précisera l'hypothèse nécessaire pour valider le calcul.*

Exercice 2.

[Le fichier-calcul disponible sur le site de la Commission Inter-IREM *Statistique et Probabilités* est *budgetpubcorrrigé.xls*]

Le tableau suivant indique les variations du chiffre d'affaires y_i d'une entreprise commerciale selon les frais de publicité x_i. (x_i et y_i sont exprimés en millions d'euros) :

x_i	7,6	8,5	9,5	10,4	11,3	12,2
y_i	198,2	224,9	228,7	247,7	266,8	274,4

1° a) - Tracer le nuage de points représentatif du tableau.

b) - Analyser le nuage quant aux 3 critères.

2°) - Si c'est envisageable, trouver une équation de la droite d'ajustement affine, par la méthode des moindres carrés ; la tracer sur le même graphique.

3°) - Donner l'interprétation concrète des coefficients *a* et *b* de l'équation.

4°) - Déduire de l'étude précédente :

a) - le chiffre d'affaires que l'entreprise réalisera si elle investit 15 millions d'euros en publicité.

b) - la somme que doit dépenser l'entreprise en publicité pour espérer un chiffre d'affaire de 800 millions d'euros.

Exercice 3. Une étude sur l'hygiène de vie : âge et force dorso-lombaire

[Le fichier-calcul disponible sur le site de la Commission Inter-IREM *Statistique et Probabilités* est *forceDLcorrigé.xls* ; voir aussi l'article d'Hubert RAYMONDAUD : *Description des séries bivariées quantitatives* (à paraître dans le volume 2 en 2005)]

L'étude suivante fait intervenir la notion naturelle de groupe (homme/femme). Cet exercice peut être traité en classe pour illustrer le cours, et servir d'exemple de situation issue de médecine et de sciences.

L'INSERM réalise des études sur le mal de dos. Cette pathologie prend une importance croissante en médecine du travail. Un des objectifs de l'étude est de connaître comment évolue la *force dorso-lombaire Y*, en fonction de *l'âge X*, chez les hommes et chez les femmes. Cela permettra de prévoir à partir de quel âge les personnes sont le plus susceptibles d'être sujettes au mal de dos.

Un extrait des résultats figure dans le tableau ci-contre : tous les résultats seront arrondis à 10^{-2} près.

Numéro	Sexe	Age (ans)	Force dorso-lombaire (dN)
1	F	44	36
2	F	38	65
3	H	20	102
4	F	23	79
5	H	44	107
6	F	62	31
7	H	62	65
8	H	32	109
9	F	29	50
10	F	50	36
11	F	53	68
12	F	47	70
13	F	35	50
14	H	35	80
15	F	32	30

Numéro	Sexe	Age (ans)	Force dorso-lombaire (dN)
16	F	26	82
17	H	59	79
18	F	41	41
19	F	20	61
20	F	56	25
21	H	56	72
22	H	41	70
23	H	47	92
24	H	50	71
25	H	29	89
26	H	53	95
27	H	26	78
28	H	38	104
29	F	59	52
30	H	23	113

1° a) - Tracer le nuage de points représentatif du tableau, en identifiant chaque point par le sexe qui lui correspond.

b) - D'après l'analyse visuelle des nuages, justifier qu'il faille trouver deux modèles : un pour les femmes et un pour les hommes.

c) - Analyser les deux nuages, quant aux trois critères.

2°) - Si c'est envisageable, trouver les équations des droites d'ajustement, pour les deux groupes, par la méthode des moindres carrés.

3° a) - Pour chaque équation, interpréter si cela a du sens, concrètement les coefficients a et b.

b) - Comparer le modèle homme et le modèle femme en faisant une phrase de commentaire comparatif des deux ordonnées à l'origine et une phrase pour les deux coefficients de régression.

4°) - Faire l'estimation de la prévision de la force dorso-lombaire pour un âge de 51 ans chez les femmes et chez les hommes.

Faire l'estimation de la prévision de l'âge pour une force dorso-lombaire de 69 dN chez les femmes et chez les hommes.

On précisera dans chacun des cas, les hypothèses nécessaires pour valider les calculs.

Exercice 4. Analyse d'une série bivariée et choix d'un modèle. Coût d'entretien et de réparation

[Le fichier-calcul disponible sur le site de la Commission Inter-IREM *Statistique et Probabilités* est *Entrepacorrigé.xls*]

L'étude du coût annuel d'entretien et de réparation C d'un équipement d'âge t, durant les cinq dernières années, a conduit à établir le tableau suivant :

Age t_i (en années)	1	2	3	4	5
Coût C_i (en euros)	2 030	2 160	2 450	2 870	3 600

1°) - Représenter graphiquement le nuage de points N(t_i , C_i). Indiquer un critère d'analyse du nuage qui ne rend pas pertinent un ajustement affine direct.

2°) - Des raisons théoriques poussent, ici, à faire le changement de variable $Z = \ln(C)$.

En vous aidant de la fiche-étude type :

a) - Représenter le nouveau nuage des points N(t_i , C_i) dans le plan muni d'un repère orthogonal. *(On pourra opérer une cassure sur l'axe des ordonnées pour optimiser la représentation !).*

b) - Justifier que l'on peut maintenant envisager un ajustement affine de ce nuage.

c) - Trouver alors, en utilisant la calculatrice, une *équation de la droite d'ajustement affine* de Z en T, par la méthode des moindres carrés. *Valeurs arrondies des coefficients à 10^{-3} près !*

d) - Évaluer la qualité de l'ajustement affine en donnant, et en commentant, le coefficient de détermination. *Valeur arrondie à 10^{-3} près !*

e) - Calculer les résidus dus au modèle $\hat{Z}$ pour chaque valeur t_i de T, et en faire une *représentation graphique.* Comment exploiter cette représentation ?

3°) - Donner alors une équation de la courbe d'ajustement exponentiel et représenter graphiquement cette courbe sur le graphique d'origine.

4°) - Estimations / Prévisions

En admettant que l'évolution du coût constaté pendant cinq ans se poursuive les années suivantes, donner une estimation du coût d'entretien et de réparation de l'équipement lorsqu'il aura 7 ans.

Exercice 5. Analyse d'une série bivariée et choix d'un modèle

[Le fichier-calcul disponible sur le site de la Commission Inter-IREM *Statistique et Probabilités* est *importexportcorrigé.xls*]

Il s'agit d'un extrait du sujet de BAC série ES 1996 (Polynésie).

Le tableau ci-après regroupe les données relatives aux importations et exportations d'un bien d'équipement en millions de francs.

Année	**1985**	**1986**	**1987**	**1988**	**1989**	**1990**	**1991**
Importations $X = x_i$	18,1	23,1	16,6	30,1	32,1	33,8	35,5
Exportations $Y = y_i$	9,4	10,9	11,9	13,7	16,5	19,7	23,1

Trouver l'ajustement qui convient le mieux pour établir la relation exprimant le montant des exportations en fonction de celui des importations ;

Toutes les droites d'ajustement demandées seront obtenues par la méthode des moindres carrés:

- Avec un ajustement exponentiel en posant $Z = \ln(Y)$.
- Avec un autre type d'ajustement en posant $Z = Y^{-2,5}$.

Activité pluridisciplinaire

Étude de la croissance d'une population microbienne

[Le fichier-calcul disponible sur le site de la Commission Inter-IREM *Statistique et Probabilités* est *Croissancemicrobienne.xls*]

Cette étude peut être menée en parallèle avec le cours de Microbiologie.

On utilise souvent l'expression *croissance exponentielle* pour qualifier un phénomène qui croît de plus en plus vite. Nous allons, dans l'étude qui suit, comprendre à l'aide d'un exemple, l'utilisation du terme *exponentielle*.

Les bactéries sont des êtres unicellulaires qui se multiplient très rapidement ; leur action biochimique est d'une importance capitale pour l'équilibre du monde vivant.

On étudie alors la croissance d'une population microbienne dans un milieu non renouvelé. Des mesures portant sur la masse de microbes, en gramme par litre ($g \times L^{-1}$), sont effectuées à divers instants t_i ; On a le tableau de valeurs suivant :

t_i (heures)	M_i ($g \times L^{-1}$)	$Z = \ln\left(\frac{M}{M_0}\right)$	$\frac{dM}{dt}$	$\frac{1}{M} \times \frac{dM}{dt}$
0	0,5			
2	0,67		0,095	
4	0,89		0,125	
6	1,18		0,167	
8	1,56		0,221	
10	2,07		0,292	
12	2,74		0,384	
14	3,63		0,504	
16	4,78		0,657	
18	6,28		0,849	
20	8,21		1,085	
22	10,65		1,36	
24	13,68		1,67	
26	17,31		1,95	
28	21,44		2,15	
30	25,75		2,11	
32	29,67		1,74	
34	32,56		1,13	
36	34,23		0,57	
38	35,00		0,24	
40	35,31		0,093	
42	35,43		0,034	

À partir des résultats figurant dans le tableau précédent :

1°) - Représenter par un nuage de points la série double (t_i, M_i) dans un repère orthogonal ; on prendra 1 cm pour 4 heures en abscisses et 1 cm pour 2 $g \times L^{-1}$ en ordonnées.

2°) - L'allure du nuage n'est pas satisfaisante pour un ajustement affine direct ! La courbe présente une croissance exponentielle les 28 premières heures, puis cette croissance diminue nettement, et la masse de microbes se stabilise.

Du moins pour les premières heures, il paraît logique de chercher une fonction f simple (affine par exemple !) telle que $M(t) = e^{f(t)}$, soit $M(t) = e^{a\,t+b}$; cela revient

donc, en faisant le changement de variable $Z = \ln(M)$, à *linéariser* le nouveau nuage représentant la série double $(t_i\ ,\ Y_i)$.

Anticipation :

Après le calcul de *a* et *b*, on obtient les valeurs estimées $\hat{Z}(t) = a\,t + b$.

Puis par la fonction réciproque du logarithme népérien, la fonction *exponentielle*, on obtient les estimations de *M*, soit $\hat{M}(t) = e^{a\,t+b} = e^{a\,t} \times e^{b}$ où, $e^{b} = \hat{M}(0)$, noté $\hat{M}_0$.

Donc par commodité, on fait plutôt le changement de variable $Z = \ln\left(\frac{M}{M_0}\right)$.

a) - Représenter par un nuage de points la série double $(t_i\ ,\ Z_i)$ dans un repère orthogonal ; on prendra 1 cm pour 4 heures en abscisses, et 1 cm pour 0,2 unités en ordonnées.

b) - Analyser le nuage et proposer un ajustement en déterminant la droite d'ajustement affine (calculs de *a*, *b* et *r*).

c) - Après avoir calculé la liste des $\hat{Z}$, calculer la liste des $\hat{M}$ (intervalles de validité à déterminer) : $\hat{Z}(t) = at + b$ pour $t \in [\ldots\ ;\ \ldots]$ et $\frac{\hat{M}(t)}{M_0} = e^{a\,t+b}$ pour $t \in [\ldots\ ;\ \ldots]$

3°) - La pente, *a*, de la droite d'ajustement, est une estimation (du moins au début) de μ, *le taux de croissance dit népérien* (vu le changement de variable) ; il indique en effet de quelle façon augmente $Z = \ln\left(\frac{M}{M_0}\right)$ en fonction de t. μ est aussi appelé *vitesse spécifique de croissance,* vu qu'il exprime une dérivée : celle de *Z(t)*, soit celle de $Z = \ln\left(\frac{M}{M_0}\right)$, soit encore celle de $\ln(M(t))$, dite dérivée logarithmique de *M*.

Cette dérivée vaut $\frac{M'(t)}{M(t)}$, ce qui s'écrit avec la notation différentielle : $\frac{1}{M} \times \frac{dM}{dt}$.

a) - Compléter alors la dernière colonne du tableau, pour trouver les valeurs prises de μ.

b) - Ces valeurs confirment-elles l'affirmation ci-dessus ?

c) - Sur le même graphique que le nuage du 2° a) et à l'aide de ces valeurs, représenter graphiquement les variations de μ en fonction de *t*.

d) - Commenter ce graphique !

Remarque :

On retrouve cette croissance exponentielle, avec le modèle discret des suites géométriques. $M_n = M_0 \times q^n = M_0 \times e^{n \ln q}$ *pour n (nombre entier d'heures) pas trop grand ; q étant le coefficient constant multiplicateur :* $\mu \approx \ln q$.

Devoir maison : Services touristiques

Relation entre revenu mensuel et budget mensuel alimentation

Cette situation de la vie économique et touristique propose un travail sur les structures.

Afin d'améliorer la qualité des services proposés dans une petite région touristique du Vaucluse, les responsables du développement du conseil général avaient effectué une enquête sur les habitudes de consommation des personnes passant visiter la région. Un extrait de cette enquête figure dans le tableau suivant. L'objectif de cette étude était de voir, pour chaque catégorie socioprofessionnelle, s'il existe une relation entre le revenu mensuel et le budget mensuel d'alimentation, puis de modéliser cette relation lorsque l'analyse montre que c'est possible.

Catégories socioprofessionnelles	Budget mensuel alimentaire	Budget mensuel d'achat de vin	Revenu mensuel	Budget annuel loisirs
Enseignant	940 €	31 €	2 459 €	1 006 €
Enseignant	773 €	28 €	2 043 €	835 €
Fonction territoriale	819 €	23 €	3 314 €	1 522 €
Ouvrier	236 €	6 €	476 €	79 €
Enseignant	1 027 €	30 €	2 702 €	1 073 €
Enseignant	236 €	23 €	2 571 €	954 €
Enseignant	966 €	23 €	2 598 €	966 €
Ouvrier	584 €	71 €	1 043 €	238 €
Fonction territoriale	686 €	22 €	3 322 €	1 490 €
Commerçant	979 €	61 €	1 679 €	537 €
Libéral	909 €	110 €	3 448 €	593 €
Fonction territoriale	788 €	23 €	3 456 €	1 523 €
Commerçant	1 101 €	70 €	1 838 €	622 €
Libéral	732 €	28 €	3 236 €	277 €
Enseignant	937 €	17 €	2 503 €	896 €
Ouvrier	525 €	50 €	1 121 €	207 €
Libéral	732 €	108 €	3 427 €	576 €
Enseignant	827 €	18 €	2 264 €	809 €
Commerçant	1 068 €	67 €	1 790 €	595 €
Commerçant	891 €	55 €	1 586 €	477 €
Ouvrier	749 €	97 €	1 402 €	312 €
Fonction territoriale	945 €	28 €	3 341 €	1 834 €
Commerçant	979 €	63 €	1 633 €	552 €
Commerçant	804 €	50 €	1 463 €	425 €
Ouvrier	595 €	60 €	1 284 €	238 €
Fonction territoriale	829 €	24 €	3 166 €	1 595 €
Libéral	890 €	127 €	3 220 €	673 €

1° a) - Tracer le nuage de points représentatif du tableau, en identifiant chaque catégorie socioprofessionnelle.

b) - Analyser le nuage ; en déduire un commentaire faisant intervenir ces catégories.

2°) - Si c'est envisageable, trouver une équation de (ou des) droite(s) d'ajustement affine, par la méthode des moindres carrés sous la forme $y = a\,x + b$.

3°) - Donner l'interprétation des coefficients *a* et *b*.

4°) - Comparer les catégories socioprofessionnelles quant à la relation entre revenu et budget d'alimentation.

Devoir surveillé (Filière Services en Milieu Rural)

[Le fichier-calcul disponible sur le site de la Commission Inter-IREM *Statistique et Probabilités* est *AnimAccueilcorRigé.xls*]

Dans les études d'un projet visant à mettre en place des structures d'accueil, d'animation et d'activités pour les jeunes de la commune, il faut PREVOIR une capacité totale d'accueil. Cette donnée est de première importance pour les décideurs et les financeurs, c'est-à-dire la commune, le département et le ministère de la Jeunesse et des Sports. Cette prévision doit être faite en fonction d'un ou plusieurs caractères, dont les valeurs chiffrées peuvent être facilement obtenues, en général à partir des données de l'INSEE.

1°) - Indiquer quels caractères (en citer quatre au maximum) vous choisiriez pour construire un modèle de prévision de la capacité totale d'accueil.

Pour simplifier le problème, on prend comme unique variable explicative la taille de la population active de la commune. On a sélectionné quelques communes ayant une structure de population se rapprochant de celle de la commune sur laquelle est préparé le projet.

Les résultats relevés figurent dans le tableau suivant :

x_i taille de la population active ($\times 10^4$)	0,978	2,013	2,967	4,021	4,989	6,104	7,012	7,987
y_i capacité totale d'accueil ($\times 10^2$)	1,02	1,23	1,68	2,59	3,54	4,80	6,79	9,97

2°) - Tracer les deux nuages de points représentant les deux séries doubles (X, Y) et $(X, \ln(Y))$ en les légendant correctement et en choisissant des échelles adaptées.

3° a) - Faire l'analyse des deux nuages.

b) - Indiquer celle des deux séries doubles que vous choisissez pour déterminer la droite d'ajustement affine et pourquoi.

4°) - Evaluer et interpréter la qualité de l'ajustement affine des deux séries et confirmer le choix fait à la question précédente.

5° a) - Trouver une équation de la droite d'ajustement affine de la série double que vous avez choisie, en donnant son domaine d'application.

b) - Si cela a du sens, interpréter ses coefficients.

c) - Tracer la droite sur le graphique légendé.

6°) - Estimer la capacité totale d'accueil à prévoir pour une commune ayant une population active de 45 000 personnes.

Devoir surveillé (Filière Industrie Agro-Alimentaire)

[Le fichier-calcul disponible sur le site de la Commission Inter-IREM *Statistique et Probabilités* est *BactériePérimcorrigé.xls*]

La qualité bactériologique des produits alimentaires est une préoccupation permanente, qui demande une vigilance de tous les instants, ce que l'actualité ne dément malheureusement pas !

Pour le nombre total de bactéries présentes dans les plats cuisinés, les normes fixées par le ministère de la santé, sont d'au maximum 70, comptées à l'aide d'une cellule de Malassez.

Afin d'établir une date limite de vente d'un plat cuisiné, un fournisseur y étudie la progression de la flore totale bactérienne (Y) à la température de conservation (2°C), en fonction du temps (X), exprimé en jours après la fabrication.

Les résultats figurent dans le tableau suivant :

x_i **nombre de jours après la fabrication**	0	1	2	3	4	5	6
y_i **nombre total de bactéries**	32,4	47,8	64,4	91,5	131,6	190,1	279,2

1°) - Tracer les deux nuages de points représentant les deux séries doubles (X, Y) et $(X, \ln(Y))$, en les légendant correctement et en choisissant des échelles adaptées.

2° a) - Faire l'analyse des deux nuages.

b) - Indiquer celle des deux séries doubles que vous choisissez pour déterminer la droite d'ajustement affine et pourquoi.

3°) - Évaluer et interpréter la qualité de l'ajustement affine des deux séries et confirmer le choix fait à la question précédente.

4° a) - Trouver une équation de la droite d'ajustement affine de la série double que vous avez choisie, en donnant son domaine d'application.

b) - Si cela a du sens, interpréter ses coefficients.

c) - Tracer la droite sur le graphique légendé.

5°) - Estimer le nombre de bactéries au bout de trois jours et demi.

6°) - Estimer la date limite de consommation de ce plat cuisiné, en utilisant la norme vue au début de l'énoncé.

Séries chronologiques

Brigitte CHAPUT

L'étude des séries chronologiques est notamment proposée par le programme de première ES dans le cadre de l'étude générale des séries statistiques. Leur spécificité temporelle conduit à des méthodes de traitement originales qui font partie de la culture statistique de base.

On appelle *série chronologique, série temporelle* ou *chronique,* une suite d'observations échelonnées dans le temps. Les intervalles entre deux observations peuvent être quelconques, mais ils sont en général de même durée : l'année, le trimestre, le mois ou le jour pour les périodicités les plus courantes. Notons $Y = (y_i)_{1 \leq i \leq n}$ une telle série statistique où i désigne le rang de l'observation y_i à l'instant t_i.

Exemple de chronique :

Le tableau suivant donne les nombres mensuels de connexions au serveur de messagerie de l'enseignement agricole depuis son ouverture en octobre 1997 jusqu'à septembre 2004. Dans cette série, nous avons n = 84 observations.

	1997-1998	**1998-1999**	**1999-2000**	**2000-2001**	**2001-2002**	**2002-2003**	**2003-2004**
Octobre	3 796	141 976	341 502	483 047	704 963	472 051	639 739
Novembre	37 423	187 369	284 915	499 734	719 495	545 235	620 169
Décembre	53 565	156 565	288 205	470 450	622 267	469 564	619 740
Janvier	75 295	219 336	360 825	602 428	822 333	647 307	817 749
Février	48 675	178 562	388 182	456 962	675 861	531 018	653 005
Mars	72 828	264 685	353 257	556 510	819 828	656 292	850 490
Avril	61 234	210 467	258 781	521 126	706 127	569 336	624 506
Mai	75 817	233 543	343 941	590 447	550 448	580 379	746 229
Juin	78 728	253 559	352 696	516 989	487 354	659 601	876 386
Juillet	43 422	142 606	221 101	378 873	300 782	398 193	440 356
Août	23 398	115 780	157 017	277 974	167 142	244 510	325 376
Septembre	92 196	274 344	424 013	593 315	463 407	712 877	936 091

Source: melagri, messagerie de l'Enseignement Agricole public - ENESAD-CNERTA - Dijon

Le traitement détaillé de cette série chronologique est proposé dans la deuxième partie de cet article et se trouve sur le site de la Commission Inter-IREM *Statistique et Probabilités* dans le classeur EXCEL *melagri.xls*.

Les grandeurs mesurées peuvent être de deux types :

- niveau ou stock : on mesure une grandeur à un instant donné (par exemple, le nombre mensuel de ventes d'un produit), c'est le cas de notre série ;
- flux : on mesure la variation d'une grandeur durant une période (par exemple, les variations mensuelles du nombre de demandeurs d'emploi).

I - Pré-traitement des données

Avant l'étude d'une série chronologique, il est parfois nécessaire de lui faire subir certains traitements. C'est le cas, en particulier, pour les séries financières lorsque l'on veut tenir compte de l'inflation ou de la dévaluation monétaire. C'est aussi le cas pour les séries chronologiques de type flux lorsque l'intervalle de temps n'est pas constant, il est alors souhaitable d'adapter les valeurs pour tenir compte des différences de durées entre deux mesures successives.

Par exemple, le nombre de jours ouvrés par mois n'étant pas constant, on peut procéder à la correction suivante, dite *des jours ouvrés* :

$$\text{valeur corrigée du mois} = \frac{(\text{valeur brute du mois}) \times (\text{nombre moyen de jours})}{\text{nombre de jours du mois}}.$$

II - Représentations graphiques

Plusieurs types de graphiques peuvent être envisagés.

1 - Diagramme cartésien

On place dans un plan rapporté à un repère orthogonal les points représentatifs des observations, de coordonnées (t_i , y_i) $1 \leq i \leq n$. C'est la représentation la plus courante, elle met en évidence les variations du phénomène étudié au cours du temps. Les points sont reliés par des segments de droite dans l'ordre croissant de leurs abscisses (voir le graphique 1).

En superposant les représentations de plusieurs périodes sur le même axe des temps, on peut aussi mettre en évidence des phénomènes saisonniers (voir le graphique 2).

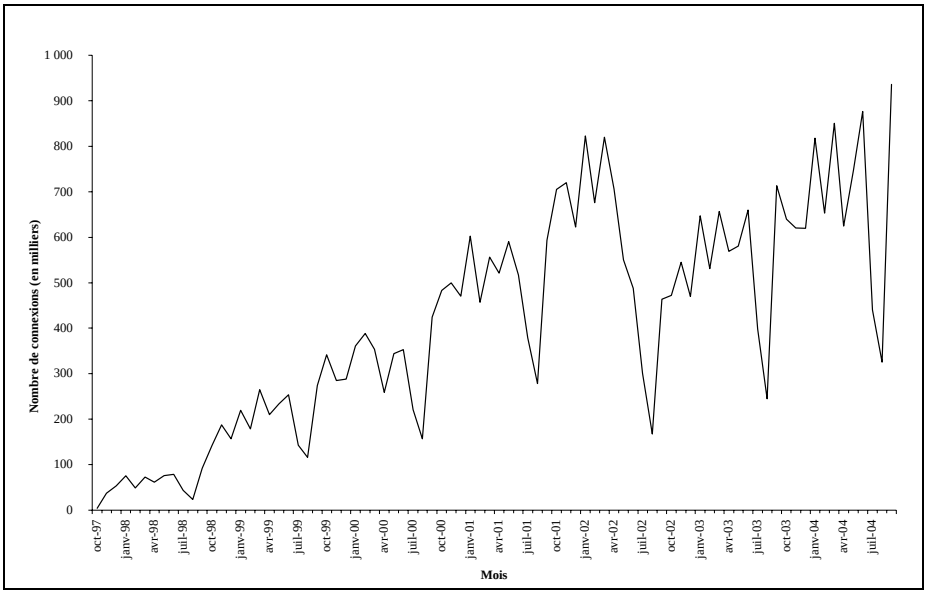
2 - Diagramme à échelle semi-logarithmique

On place dans un plan rapporté à un repère orthogonal les points de coordonnées $(t_i , \ln(y_i))$ ou $(t_i , \log(y_i))$. Comme pour la représentation précédente, les points sont reliés par des segments de droite dans l'ordre croissant de leurs abscisses. (On peut utiliser du papier semi-logarithmique dont une des échelles de graduation est logarithmique décimale et l'autre régulière.)

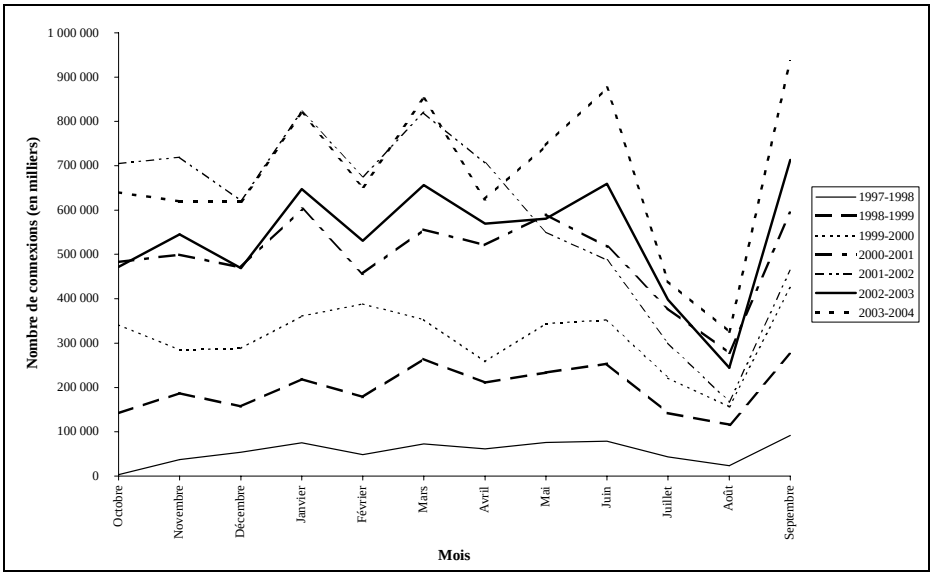
On utilise un diagramme à échelle semi-logarithmique quand :

- les valeurs observées présentent une si grande variabilité qu'une échelle régulière ne permet pas une représentation révélatrice ;

– on veut mettre en évidence des propriétés du phénomène qui n'apparaissent pas aussi clairement en échelle arithmétique. Notamment, une échelle logarithmique permet de lire directement un taux de croissance relative.



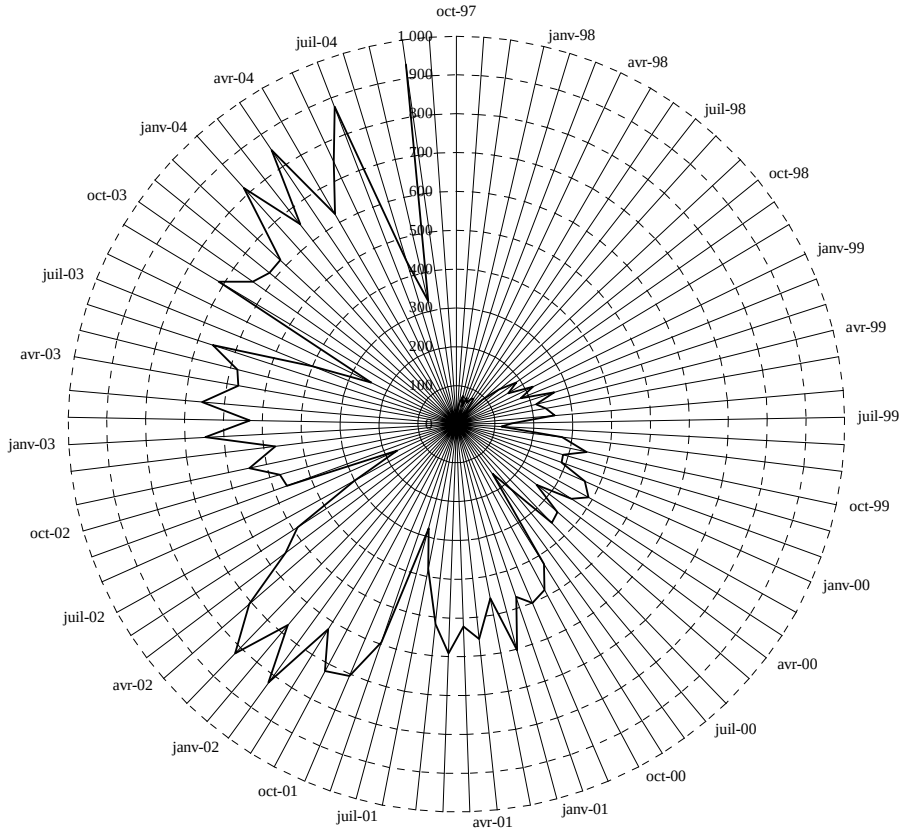
Graphique 1 : représentation de la série de l'exemple sur un diagramme cartésien.



Graphique 2 : superposition des représentations des sous-séries annuelles.

3 - Diagramme polaire

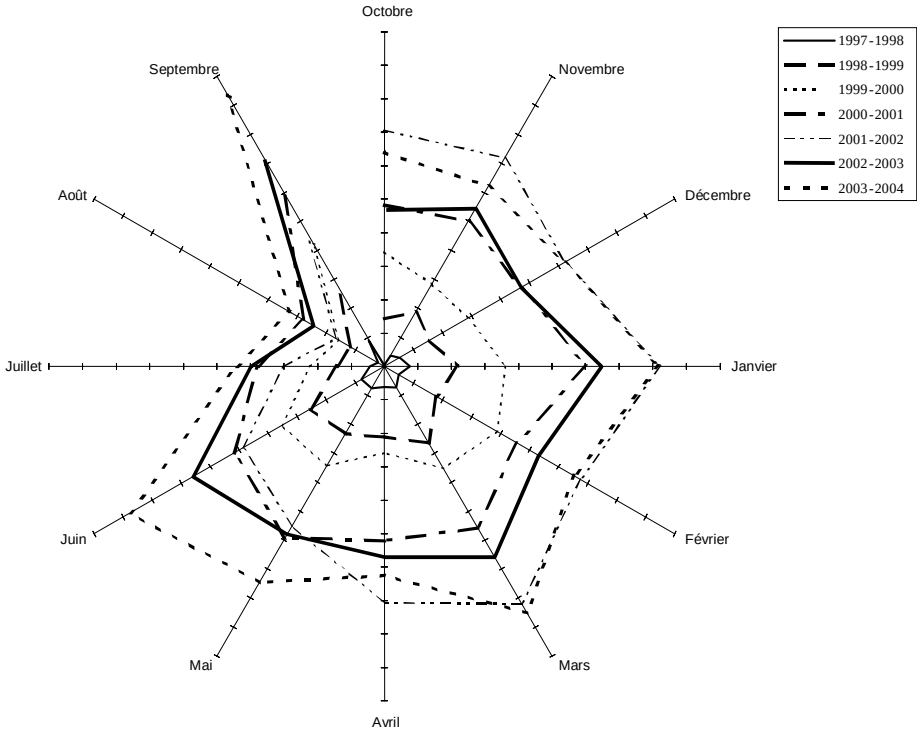
On fixe un point O, le pôle, et une demi-droite d'origine O, l'origine des angles. Chaque valeur de la chronique est représentée par un point M tel que la distance OM est proportionnelle à la valeur observée, et l'angle entre la demi-droite origine des angles et [OM) est proportionnel au temps écoulé. On relie ensuite les points successifs par des segments de droite dans l'ordre croissant des mesures d'angle.



Graphique 3 : représentation de la série de l'exemple sur un diagramme polaire.

Ces graphiques nous permettent une première remarque : dans la série que nous étudions, le diagramme cartésien et le diagramme polaire mettent en évidence une rupture dans la croissance des connexions à partir d’avril 2002. Cette rupture est expliquée dans l’étude détaillée du paragraphe V, page 120.

La superposition des représentations polaires correspondant à des périodes successives permet de mettre en évidence les phénomènes saisonniers et la croissance ou la décroissance du phénomène.



Graphique 4 : superposition des représentations polaires des sous-séries annuelles

III - Mouvements caractéristiques d'une série chronologique

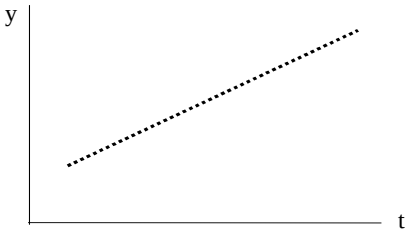
Les séries chronologiques combinent un certain nombre de mouvements et de variations caractéristiques dont certains se manifestent à des degrés variés. L'analyse de ces mécanismes est importante, en particulier pour prévoir les variations futures.

Dans l'évolution d'une série chronologique Y, on peut distinguer quatre composantes, additives ou multiplicatives selon les modèles adoptés : la tendance à long terme T, les mouvements cycliques C, les variations saisonnières S et les composantes aléatoires A.

1 - La tendance

La *tendance* ou *trend* T(t) représente le mouvement profond de l'évolution à long terme du phénomène, contribuant aux variations de la série Y. On la représente par une courbe de tendance en tirets.

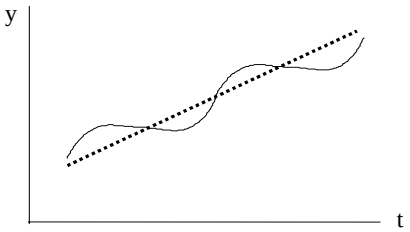
Exemples : diminution de la population active agricole ; croissance de la production industrielle ou de la consommation d'électricité.



Graphique 5 : tendance indiquant une croissance globale régulière

2 - Les mouvements cycliques

On peut identifier des *mouvements cycliques* C(t) qui fluctuent autour de la tendance à long terme et qui sont liés aux variations conjoncturelles (par exemple, à la succession des phases du cycle économique : prospérité, crise, dépression, reprise).



Graphique 6

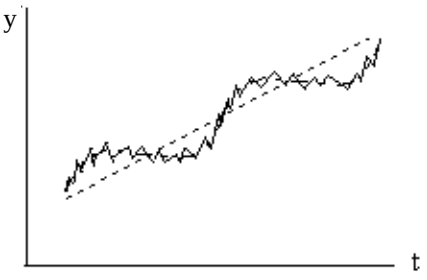
En général de grande périodicité, cette contribution aux variations d'une série chronologique Y ne peut être mise en évidence que sur des séries assez longues ; on l'ignore quand les données ne remontent pas suffisamment dans le temps.

Exemple : Le taux annuel de progression de la production industrielle a pu atteindre 12 % en période de prospérité et s'annuler en période de dépression, alors que le taux moyen représentant la tendance à long terme est d'environ 6,5 %.

3 - Les variations saisonnières

On appelle *variations saisonnières* ou *saisonnalité*, des fluctuations périodiques S(t) de même type que les mouvements cycliques mais de période plus courte.

Les variations saisonnières peuvent avoir une période journalière (trafic horaire du métro), hebdomadaire (nombre d'heures travaillées par jour) ou encore annuelle (indice mensuel de la production industrielle, chiffre d'affaires des grands magasins).



Graphique 7 : saisonnalité sur un mouvement cyclique

Les variations saisonnières ont de multiples causes : cycle des saisons, dispositions réglementaires dont les effets se produisent à date fixe. Voici quelques exemples :

- les congés : les congés annuels se traduisent chaque été par un ralentissement sensible de l'activité et une diminution des principales grandeurs économiques ;
- l'inégalité des différents mois (nombre de jours, nombre de fêtes mobiles...) ;
- les facteurs climatiques qui influent sur l'activité de l'industrie du bâtiment, sur la consommation d'électricité... ;
- la périodicité de l'offre et de la demande de certains produits (rythme saisonnier de la production agricole, ventes de fin d'année, demande d'automobiles au printemps...).

4 - Les composantes aléatoires

Les *composantes aléatoires* ou *irrégulières* prennent en compte les aspects aléatoires des variations de la série chronologique. On espère que cette contribution imprévisible, notée A(t) est assez faible par rapport aux autres valeurs T(t), C(t) et S(t). On considère que ces composantes comprennent aussi tout ce qui n'a pas été pris en compte par les autres composantes du modèle. Elles font intervenir des causes conjoncturelles ou accidentelles, et rendent compte de phénomènes particuliers, limités dans le temps (grèves, actions volontaristes ou publicitaires...).

IV - Quelques modèles

L'étude des séries chronologiques permet de traiter un certain nombre de questions (prévision, suppression de la tendance, correction des variations saisonnières...).

L'analyse d'une série chronologique consiste à faire une description des mouvements la composant. On émet des hypothèses quant à leur influence sur la série chronologique Y étudiée. Deux modèles sont choisis couramment :

- le *modèle additif* où l'on considère que les effets des différents mouvements s'ajoutent : $Y(t) = T(t) + C(t) + S(t) + A(t)$;
- le modèle *multiplicatif* où l'on préfère considérer ces effets comme multiplicatifs : $Y(t) = T(t)\, C(t)\, S(t)\, A(t)$.

Après cette brève introduction, nous allons traiter plus complètement l'exemple introduit au début sur les connexions mensuelles au serveur de l'enseignement agricole.

V - Analyse de la série donnée en exemple

1 - Détermination de la tendance

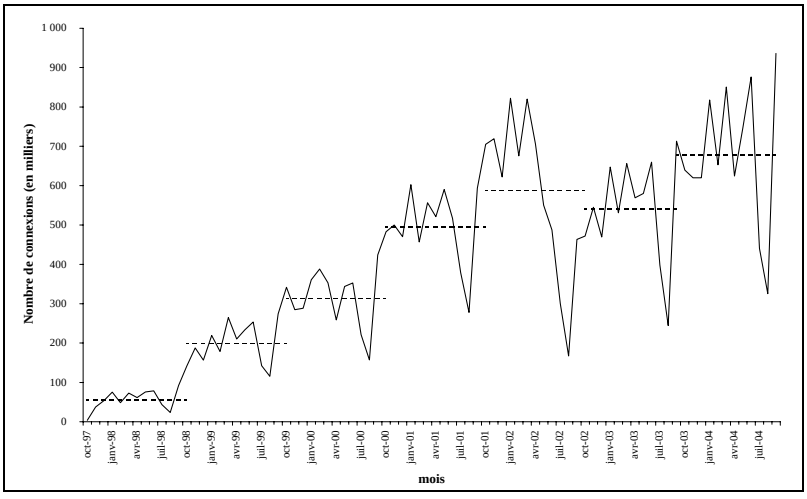
On peut déterminer les valeurs T(t) de la tendance par diverses méthodes d'ajustement.

- **Ajustement par les moyennes annuelles**

Pour chaque année, on donne la moyenne des observations de l'année :

Année	**1997-98**	**1998-99**	**1999-00**	**2000-01**	**2001-02**	**2002-03**	**2003-04**
Moyenne annuelle	55 531,4	198 232,7	314 536,3	495 654,6	586 667,3	540 530,3	679 153,0

Le graphique 8 suivant associe aux données leurs moyennes annuelles représentées par des segments en pointillés. La tendance est alors représentée par une fonction en escaliers :



Ce graphique montre la rupture dans la tendance, déjà mise en évidence par les diagrammes cartésien et polaire des pages 113 et 114.

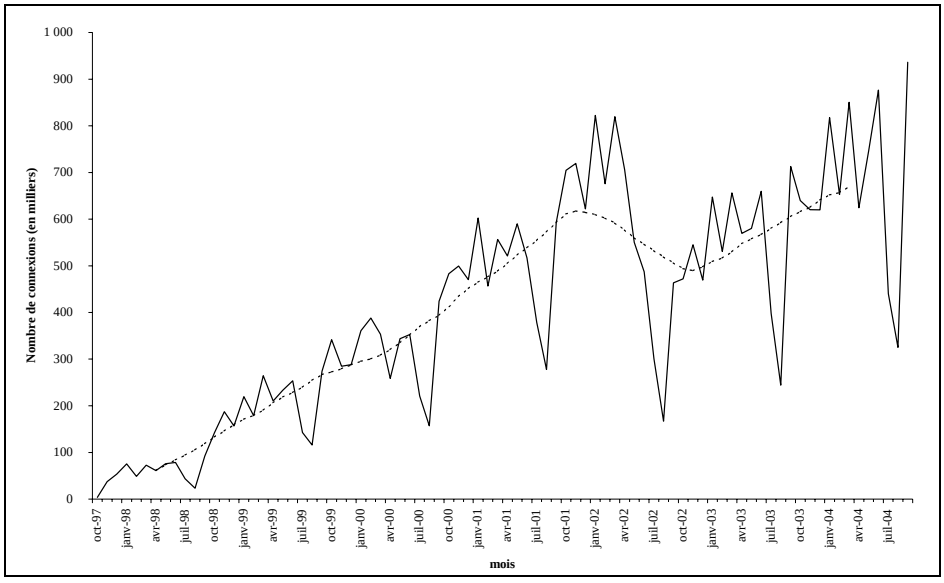
- **Ajustement par la méthode des moyennes mobiles**

La méthode des moyennes mobiles consiste à faire correspondre à chaque observation la moyenne d'ordre p des données qui l'encadrent. Dans notre exemple, à chaque mois i on associe la valeur m_i de la moyenne centrée d'ordre 12, ainsi calculée : $m_i = \dfrac{y_{i-6} + 2(y_{i-5} + y_{i-4} + \ldots + y_{i+4} + y_{i+5}) + y_{i+6}}{24}$.

	1997-1998	1998-1999	1999-2000	2000-2001	2001-2002	2002-2003	2003-2004
Octobre		133 158,7	272 611,8	413 154,4	611 331,0	494 027,6	616 114,4
Novembre		145 948,7	279 224,8	434 356,5	617 372,7	489 575,1	625 323,6
Décembre		159 805,2	287 955,5	451 473,1	614 471,3	497 999,2	641 266,7
Janvier		171 222,5	295 356,8	464 892,5	609 982,7	509 235,0	652 056,2
Février		179 204,4	300 345,6	476 506,2	602 110,9	516 517,4	657 182,4
Mars		190 643,2	308 300,0	488 600,3	592 080,1	530 135,7	669 852,4
Avril	61 288,9	206 546,3	320 434,0	504 901,1	576 962,6	547 517,3	
Mai	73 294,2	218 924,3	335 282,5	523 304,3	559 997,1	557 626,5	
Juin	83 833,6	228 473,7	351 826,8	538 786,7	546 373,6	567 006,1	
Juillet	94 127,0	239 854,0	369 487,1	554 275,1	532 718,3	580 365,2	
Août	105 540,6	254 483,6	382 419,8	572 558,6	519 390,4	592 549,7	
Septembre	118 946,6	266 908,3	393 754,5	592 651,0	506 541,3	605 724,1	

Tableau des moyennes mobiles centrées d'ordre 12

Sur le graphique 9 suivant, la tendance est représentée par la courbe en pointillés des moyennes mobiles ainsi calculées.



On voit que la variabilité de la série des moyennes mobiles est plus faible que celle des données. On obtient ainsi un *lissage* de la série chronologique et une meilleure vision du mouvement général de ses valeurs, ayant réduit les effets des fluctuations ponctuelles, aléatoires ou accidentelles.

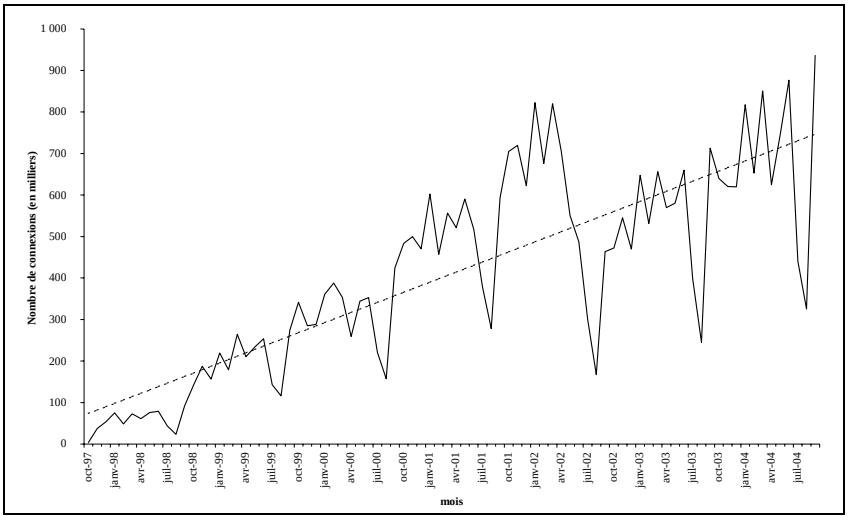
Cette méthode présente cependant plusieurs inconvénients :

- on perd les données de début et de fin de série, l'usage de la méthode est donc réservé aux séries chronologiques suffisamment longues ;
- les moyennes mobiles peuvent donner lieu à des cycles et des mouvements qui n'étaient pas présents dans les données d'origine ;
- une valeur « aberrante » accidentelle affecte toutes les moyennes mobiles voisines.

- **Ajustement par la méthode des moindres carrés**

On ajuste les points de la série par la droite des moindres carrés (voir l'article de Stéphan MANGANELLI, p. 83) qui peut ainsi représenter la tendance.

Dans notre exemple, le mois étant pris pour unité de temps à partir de début octobre 97 et les connexions comptées en milliers, une équation approchée de cette droite est $y = 8{,}1\,t + 65{,}4$.



Graphique 10 : ajustement de la série par la méthode des moindres carrés

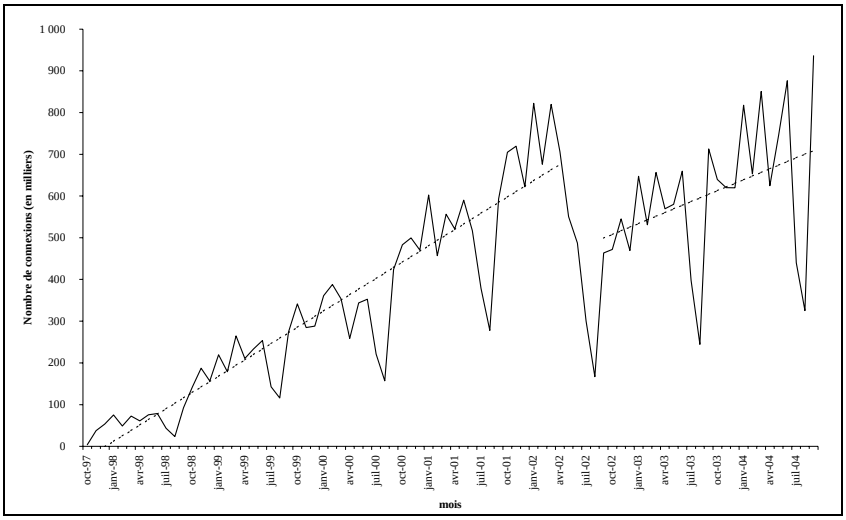
Les méthodes précédentes sont simples d'utilisation, mais elles peuvent conduire à des résultats peu significatifs si elles sont utilisées sans discernement, le traitement de notre exemple avec la méthode des moindres carrés en est une belle illustration. En effet, cette méthode ne met pas en évidence la rupture de tendance constatée avec les autres méthodes. Interrogée à ce propos, la personne gestionnaire du serveur nous en a indiqué la raison : la diminution systématique du nombre de connexions à partir de juin 2002, est consécutive à une saturation du serveur qui a commencé en avril 2002. Rencontrant des difficultés de connexions,

les personnes consultant le serveur se sont découragées et beaucoup ont abandonné leurs tentatives. Le serveur a été adapté en septembre 2002 et la reprise des connexions s'est faite ensuite progressivement.

Pour mieux rendre compte de la situation et avoir une meilleure image de la tendance, il est plus significatif de séparer l'ajustement en deux parties.

La droite d'ajustement des moindres carrés correspondant à la première période (avant avril 2002) a pour équation approchée $y = 13\,t - 40{,}1$, celle de la seconde période (après septembre 2002) est $y = 8{,}7\,t - 25{,}7$.

Le graphique 11 ci-dessous montre bien cette rupture de tendance dont le calcul n'a pas de sens entre mai et septembre.



2 - Estimation des variations saisonnières, indices saisonniers

a) - Détermination de la périodicité des variations saisonnières

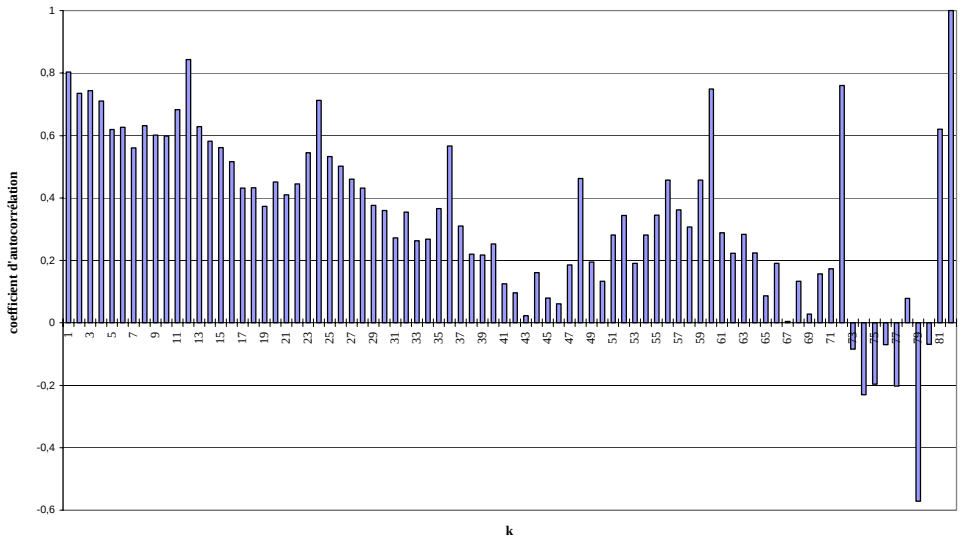
Lorsque la saisonnalité n'apparaît pas nettement sur le graphique cartésien, on utilise un graphique appelé *corrélogramme*. C'est le diagramme en bâtons qui représente la *fonction d'autocorrélation* : à $k \in \mathbb{N}^*$, elle associe $\rho(k)$, le coefficient de corrélation linéaire entre la série chronologique (y_i) et la série (y_{i+k}) obtenue en décalant la série initiale de k unités de temps.

$\rho(k) = \dfrac{\mathrm{Cov}((y_i)\,,\,(y_{i+k}))}{\sigma(y_i)\,\sigma(y_{i+k})}$ s'appelle le *coefficient d'autocorrélation d'ordre k.*

Lorsque tous les $\rho(p \times j)$ sont voisins de 1, les séries (y_i) et (y_{i+pj}) sont voisines et la série chronologique présente une saisonnalité d'ordre p.

Le corrélogramme obtenu pour la série des connexions au serveur de l'enseignement agricole suggère une périodicité de 12 mois correspondant à une saisonnalité annuelle.

Dans le graphique 12 ci-dessous, on remarque cette périodicité sur les pics de valeurs du coefficient d'autocorrélation, ainsi que sur l'allure générale des périodes de 12 mois reproduisant en décroissance globale les variations de ce coefficient. Cette décroissance est due au fait que les deux séries (y_i) et (y_{i+k}) sont de moins en moins corrélées quand le décalage k augmente.



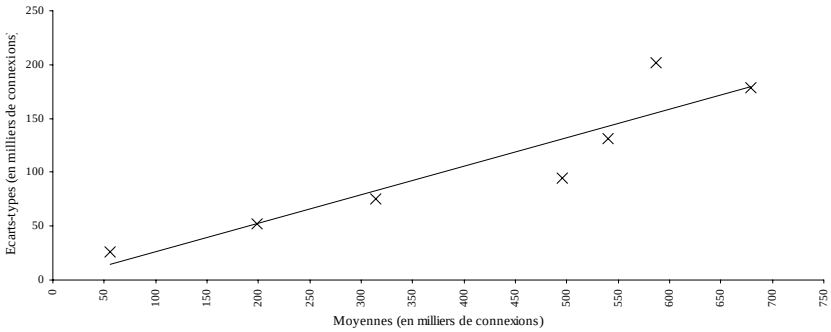
b) - Détermination du modèle

Dans un modèle multiplicatif, l'effet saisonnier et l'effet aléatoire sont amplifiés (ou diminués) proportionnellement à la croissance (ou décroissance) de la tendance. Ceci peut se voir sur la représentation en graphique cartésien de la série chronologique, en observant ses fluctuations autour de la tendance.

Pour le mettre en évidence, on utilise un graphique appelé *diagramme mu-sigma*. Lorsque la saisonnalité est d'ordre p, on calcule les moyennes μ et les écarts-types σ des p premières valeurs de la série chronologique, puis des p suivantes et ainsi de suite ; on porte alors sur un diagramme cartésien les points d'abscisses μ et d'ordonnées σ, il y a alors autant de points que d'intervalles d'étude de longueur p.

Dans le cas du modèle additif, ces points auront tendance à être alignés horizontalement (les fluctuations restant constantes), tandis que dans le modèle multiplicatif, ils s'aligneront à peu près sur une droite oblique.

Le diagramme mu-sigma de notre exemple fait l'objet du graphique 13 suivant (l'étude est réalisée sur 7 ans et la saisonnalité est annuelle, il y a 7 points).



La droite d'ajustement du nuage des (μ, σ) est oblique, ce qui nous suggère d'adopter un modèle multiplicatif pour analyser les variations de cette série, afin de séparer les contributions de la tendance, des variations cycliques et saisonnières et des fluctuations aléatoires. Les valeurs observées y_i de la série Y seront donc considérées comme des produits : $y_i = T(t_i)\, C(t_i)\, S(t_i)\, A(t_i)$.

La démarche d'analyse pour un modèle additif serait analogue à celle qui est présentée maintenant pour un modèle multiplicatif.

c) - Estimation des variations saisonnières

On suppose que la saisonnalité est d'ordre p et que l'étude de la série chronologique se fait sur α périodes. Dans notre exemple, on considère qu'il n'y a pas d'influence cyclique (C(t) = 1), on a $p = 12$ et $\alpha = 7$. De plus, le graphique 11 de la page 121 nous permet de considérer que les variations saisonnières et aléatoires ne sont pas notablement affectées par la rupture de tendance soulignée page 120. Nous retenons donc cette hypothèse pour déterminer les variations saisonnières à partir de l'ensemble des données.

Les quotients, exprimés en pourcentages, $z_i = \frac{y_i}{T(t_i)}$ des valeurs observées de la série chronologique par les valeurs de la tendance retenue, rendent compte des composantes saisonnières, cycliques et aléatoires du phénomène étudié, mais leur série n'est pas périodique en général.

On calcule alors les p valeurs X(k), moyennes arithmétiques des z_k, z_{k+p}, ..., $z_{k+(\alpha-1)p}$, pour k entier compris entre 1 et p. Ce calcul a pour but de réduire les effets des fluctuations accidentelles qui ne se reproduisent pas d'année en année. Les X(k) représentent donc la contribution des variations cycliques et saisonnières.

Dans l'exemple, les X(k) ($1 \le k \le 12$) sont les valeurs moyennes sur les 7 années d'étude des pourcentages des valeurs observées rapportées à la tendance générale, pour chacun des mois de l'année.

En rapportant ces X(k) à leur valeur moyenne, on obtient la composante mensuelle saisonnière donnée sur une période par $S(t_k) = \frac{X(k) \times p}{X(1)+X(2)+...+X(p)}$; cette quantité exprimée en pourcentage, porte le nom d'*indice saisonnier* noté I(k).

Les indices saisonniers I(k) sont donc sans dimension et ont pour moyenne 100 %. Dans notre exemple, ils rendent compte uniquement des composantes saisonnières de la série étudiée, car nous avons supposé l'absence d'effet cyclique.

Il y a autant de méthodes de calcul des indices saisonniers que de méthodes de détermination de la tendance, mais toutes reposent sur le même principe, que nous résumons ici dans le cas du modèle multiplicatif de l'exemple :

- On détermine la tendance des observations, représentée par une fonction T(t). Pour notre exemple, nous avons retenu l'ajustement par les moindres carrés, qui, comme on l'a vu, n'a pas de signification de mai à août 2002.
- On exprime chaque donnée y_i en pourcentage de la valeur correspondante de la tendance $z_i = \frac{y_i}{T(t_i)}$. Cela donne le tableau numérique suivant :

	1997-1998	**1998-1999**	**1999-2000**	**2000-2001**	**2001-2002**	**2002-2003**	**2003-2004**
Octobre		110,0 %	119,7 %	109,4 %	117,9 %	93,0 %	104,4 %
Novembre		131,9 %	95,5 %	110,0 %	117,8 %	105,6 %	99,8 %
Décembre		100,9 %	92,6 %	100,6 %	99,8 %	89,4 %	98,3 %
Janvier		130,5 %	111,3 %	125,4 %	129,1 %	121,2 %	128,0 %
Février		98,6 %	115,1 %	92,6 %	104,0 %	97,8 %	100,8 %
Mars		136,3 %	100,8 %	109,9 %	123,7 %	119,0 %	129,6 %
Avril	120,1 %	101,6 %	71,2 %	100,3 %	104,5 %	101,6 %	93,9 %
Mai	118,5 %	106,1 %	91,4 %	110,9 %		102,0 %	110,7 %
Juin	102,2 %	108,7 %	90,6 %	94,8 %		114,2 %	128,4 %
Juillet	48,2 %	57,9 %	54,9 %	67,8 %		67,9 %	63,7 %
Août	22,7 %	44,7 %	37,8 %	48,6 %		41,1 %	46,5 %
Septembre	79,4 %	100,8 %	99,0 %	101,5 %	92,9 %	118,0 %	132,0 %

Tableau des z_i

Par exemple, pour la valeur 119,7 % du mois d'octobre 1999 (mois d'abscisse 25), on divise la valeur observée 341,5 par la valeur $T(t_{25}) = 285$ tirée de l'équation de la droite des moindres carrés $y = 13\,t - 40{,}1$.

- On calcule ensuite pour chaque mois les composantes saisonnières X(k), moyennes des pourcentages z_i obtenus sur les mêmes mois de chaque année.
- On rapporte ces composantes X(k) à leurs moyennes de façon à obtenir les indices saisonniers I(k) relatifs à chacun des mois.

Pour notre exemple, on obtient le tableau des composantes saisonnières et des indices saisonniers suivant :

	Octobre	Novembre	Décembre	Janvier	Février	Mars
X(k)	109,1 %	110,1 %	96,9 %	124,2 %	101,5 %	119,9 %
Indices saisonniers I(k)	111,2 %	112,2 %	98,8 %	126,6 %	103,4 %	122,2 %

	Avril	Mai	Juin	Juillet	Août	Septembre
X(k)	99,0 %	106,6 %	106,5 %	60,1 %	40,2 %	103,4 %
Indices saisonniers I(k)	100,9 %	108,6 %	108,5 %	61,2 %	41,0 %	105,3 %

La composante saisonnière des mois de mai, moyenne X(8) des 6 valeurs de z_i associées aux mois de mai retenus, est égale à 106,6 %.

La moyenne des X(k) vaut 98,1 % et les indices saisonniers I(k) valent $\frac{X(k)}{98,1}$.

d) - Désaisonnalisation des données

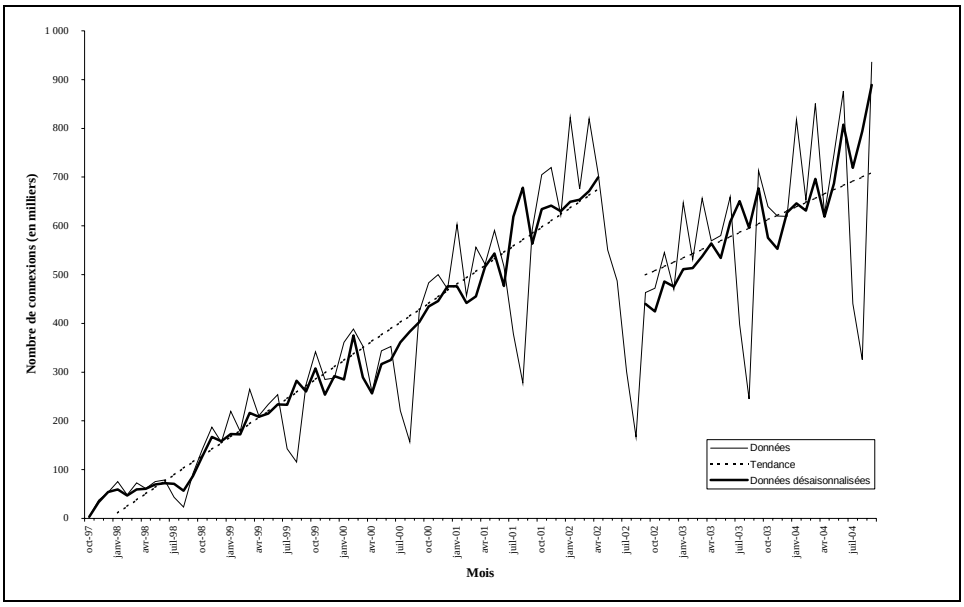
Quand on divise les y_i par les $S(t_i)$ associés (ou dans un modèle additif, on retranche les $S(t_i)$ aux y_i), on dit qu'on *ajuste les données suivant les variations saisonnières*. Les valeurs obtenues prennent alors en compte la tendance, les mouvements cycliques et irréguliers, selon le schéma $\frac{Y}{S} = T\ C\ A$.

	1997-1998	1998-1999	1999-2000	2000-2001	2001-2002	2002-2003	2003-2004
Octobre		127 719,6	307 210,5	434 542,4	634 175,0	424 650,6	575 500,4
Novembre		167 009,2	253 955,6	445 432,0	641 313,4	485 988,8	552 780,4
Décembre		158 470,5	291 712,7	476 175,8	629 840,5	475 279,0	627 282,7
Janvier		173 227,6	284 973,1	475 786,8	649 463,8	511 231,4	645 843,4
Février		172 643,2	375 314,9	441 815,1	653 458,3	513 416,4	631 359,9
Mars		216 645,7	289 142,3	455 505,7	671 032,5	537 177,6	696 129,5
Avril	60 672,8	208 538,2	256 409,4	516 350,2	699 655,7	564 118,4	618 782,7
Mai	69 798,4	215 003,5	316 637,7	543 575,2		534 306,5	686 990,7
Juin	72 555,0	233 677,7	325 041,5	476 452,5		607 882,5	807 669,6
Juillet	70 913,2	232 892,4	361 084,0	618 744,3		650 296,1	719 153,2
Août	57 077,8	282 437,1	383 031,8	678 097,9		596 464,8	793 731,7
Septembre	87 515,5	260 416,5	402 487,3	563 194,4	439 881,4	676 686,7	888 568,9

Tableau des données désaisonnalisées

Par exemple, la valeur 307 210,5 associée à octobre 1999 est égale à $\frac{341\,502}{0,1112}$.

Le graphique 14 suivant présente ces données désaisonnalisées :



3 - Détermination de la composante cyclique

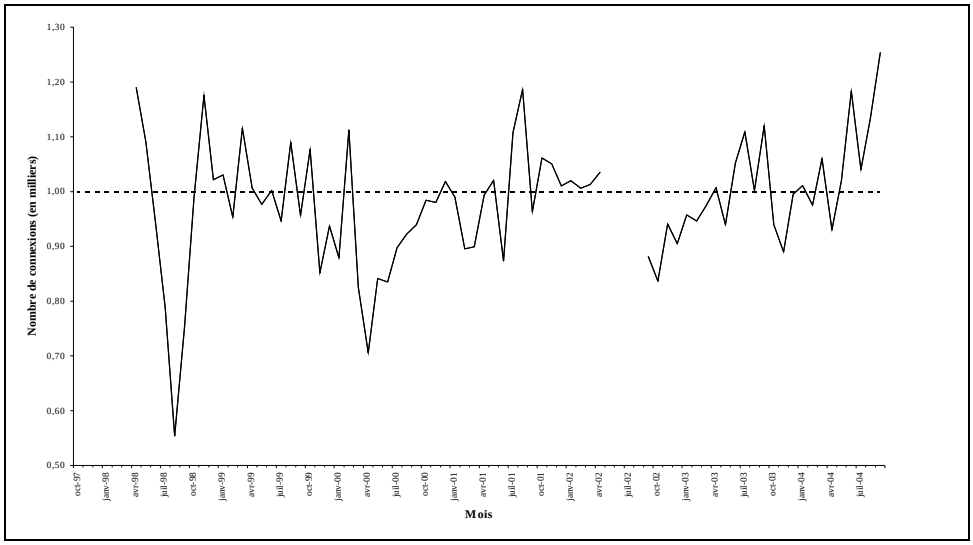
La mise en évidence des variations cycliques C(t) se fait de manière analogue à celle des variations saisonnières à partir des données désaisonnalisées.

Dans notre exemple, il n'est pas raisonnable d'étudier ce mouvement en raison de la courte durée d'étude de la série chronologique, nous avons donc fait l'hypothèse de le négliger en prenant C(t) = 1.

4 - Les composantes aléatoires

Les composantes aléatoires sont obtenues en divisant (ou en soustrayant dans un modèle additif) les données désaisonnalisées par la tendance retenue, selon le schéma $\frac{Y}{S\,T} = A$.

Dans le cas de l'exemple, ces valeurs sont proches de 1 et réparties aléatoirement autour de 1, leur impact sur les valeurs observées se confirme donc comme minime.



Graphique 15 : variations aléatoires

VI - Prévision

Le problème de la prévision repose sur deux questions fondamentales :

– Quelle(s) est (sont) la (les) grandeur(s) à prévoir ?

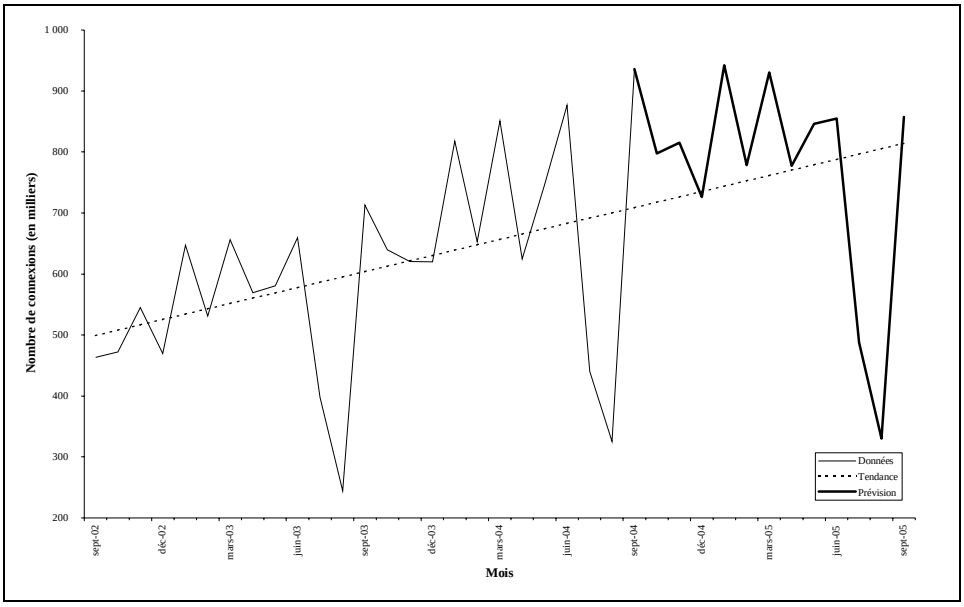
– Pour quoi faire ?

Le problème se pose alors dans les termes suivants : « On connaît y_1, y_2, ..., y_n et on cherche à proposer des valeurs pour y_{n+1}, y_{n+2}, ..., y_{n+h}, h étant l'*horizon* de la prévision ».

Dans notre exemple, on ne peut faire raisonnablement de prévision qu'à partir des données postérieures à septembre 2002. Le petit nombre de ces données ne permet pas une grande précision à long terme. On se base sur la tendance T(t) déterminée par la deuxième droite d'ajustement donnée page 121, modulée par les variations saisonnières en multipliant T(t) par S(t), négligeant une composante cyclique éventuelle C(t), ainsi que les variations aléatoires qui se sont révélées également négligeables jusqu'à présent. On obtient ainsi une prévision pour l'année scolaire 2004-2005 : $y_{prévu} = [8{,}75\,(84 + k) - 25{,}7] \times I(k)$, où k varie de 1 à 12.

Mois	**Oct. 2004**	**Nov. 2004**	**Déc. 2004**	**Janv. 2005**	**Fév. 2005**	**Mars 2005**
Prévision	797 791,9	814 986,4	726 334,5	941 931,6	778 470,5	930 249,1

Mois	**Avr. 2005**	**Mai 2005**	**Juin 2005**	**Juil. 2005**	**Août 2005**	**Sept. 2005**
Prévision	777 282,0	846 068,5	854 663,8	487 654,1	330 053,7	857 416,7



Graphique 16 : prévisions du nombre de connexions en 2004-2005

Les idées précédentes permettent de résoudre le problème important de la prévision des séries temporelles. Cependant, il faut prendre garde que le traitement mathématique des données ne résout pas tous les problèmes. L'analyse mathématique est néanmoins appréciable si le statisticien y associe du bon sens, de l'ingéniosité et un bon jugement.

Bibliographie

CHAUVAT Gérard, RÉAU Jean-Philippe, *Statistique descriptive*, Collection Les Fondamentaux, Hachette, 1995

CHAUVAT Gérard, RÉAU Jean-Philippe, *Statistiques descriptives. Exercices et corrigés*, Collection Cursus, Armand Colin, 1992

SCHLACTHER Didier, *De l'analyse à la prévision*, Collection Statistique pour les sciences économiques et sociales, Ellipses, 1986

SPIEGEL Murray R., *Statistique. Cours et problèmes*, Série Schaum, McGraw Hill (Ediscience international), 1993

Derrière la statistique, la géométrie[1]

Jean Claude GIRARD et Brigitte CHAPUT

Les statistiques prennent généralement pour les élèves (et aussi pour leurs professeurs !) la forme de calculs ou d'applications de formules dont la justification et le rapport avec les mathématiques n'apparaissent pas clairement.

Cet article a pour objectif de montrer que la statistique est partie intégrante des mathématiques en présentant quelques notions classiques dans un cadre géométrique. Il n'est qu'une prise de contact très limitée avec le vaste champ de l'analyse exploratoire. Insuffisante en elle-même pour assimiler réellement les résultats qui sont présentés ici, cette introduction aux outils modernes de la statistique descriptive appelle un approfondissement éclairé de nombreux exemples que l'on trouvera dans la bibliographie, notamment dans l'ouvrage de référence de G. SAPORTA *Probabilités, statistique et analyse des données.*

I - Présentation du cadre

Une série statistique X définie sur une population de n individus, peut être assimilée au vecteur $(x_1 , x_2 , ... , x_n)$ de $\mathbb{R}^n$ où x_i est la valeur de X observée sur le $i^{ème}$ individu.

On munit $\mathbb{R}^n$ de la métrique définie par la matrice symétrique, définie, positive

$$M = \begin{pmatrix} \frac{1}{n} & 0 & - & - & 0 \\ 0 & \frac{1}{n} & - & - & 0 \\ 0 & 0 & - & - & - \\ - & - & - & - & - \\ 0 & 0 & - & - & \frac{1}{n} \end{pmatrix} = \frac{1}{n} I_n$$

où I_n désigne la matrice unité d'ordre n.

[1] On trouvera une autre présentation de ce thème *Statistique et géométrie* dans le chapitre 4 *Statistique euclidienne* de **[1]**, pages 69 à 80.

Les définitions classiques dans les espaces euclidiens donnent pour la métrique choisie :

Produit scalaire de deux vecteurs

$$\langle u , v\rangle = {}^t u\, M\, v = \frac{1}{n} x_1 y_1 + \ldots + \frac{1}{n} x_n y_n = \frac{1}{n} \sum_{i=1}^{n} x_i y_i$$

Norme d'un vecteur

$$\|u\|^2 = \langle u , u\rangle = \frac{1}{n} x_1^2 + \ldots + \frac{1}{n} x_n^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2$$

Cosinus de l'angle de deux vecteurs non nuls

$$\cos(u , v) = \frac{\langle u , v\rangle}{\|u\|\,\|v\|} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2}}$$

où $u = \begin{pmatrix} x_1 \\ x_2 \\ - \\ - \\ x_n \end{pmatrix}$ et $v = \begin{pmatrix} y_1 \\ y_2 \\ - \\ - \\ y_n \end{pmatrix}$

Projeté de X sur Y (Y non nul)

$$\mathrm{proj}_v(u) = \frac{\langle u , v\rangle}{\|v\|^2} v$$

II - Projection orthogonale sur un axe

1 - Moyenne, écart-type

Problème

Étant donnée une série statistique $X = (x_1 , x_2 , \ldots , x_n)$, considérée comme un vecteur de $\mathbb{R}^n$, on cherche à la remplacer par un vecteur $(a , a , \ldots , a)$ $(a \in \mathbb{R})$ qui soit le plus proche possible de X pour la distance définie par la métrique considérée[2].

Les vecteurs de la forme $(z , z , \ldots , z)$ forment la droite vectorielle Δ engendrée par le vecteur $\mathbb{1} = (1 , 1 , \ldots , 1)$. Cette droite vectorielle est la n-sectrice des axes. Le vecteur le plus proche de X est le projeté orthogonal de X sur Δ.

Comme $\|\mathbb{1}\| = 1$, $\mathrm{proj}_\Delta(X) = \langle X , \mathbb{1}\rangle\, \mathbb{1} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i \times 1\right) \mathbb{1} = \bar{x}\, \mathbb{1} = \overline{X}$.

Le vecteur $\overline{X}$ a toutes ses composantes égales à la moyenne des données observées de la série statistique X.

[2] Comme la matrice M est proportionnelle à la matrice unité I_n, un vecteur $(a , a , \ldots , a)$ solution est donc aussi le plus proche de X pour la métrique euclidienne.

Illustration[3]

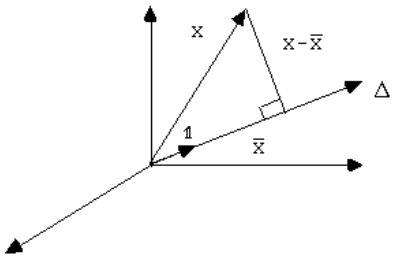


Figure 1

Par conséquent, $\overline{X} = (\bar{x}, \bar{x}, \ldots, \bar{x})$, le vecteur de Δ le plus proche de X, est aussi le vecteur de Δ qui minimise $||X - a\mathbb{1}||$ quand $a \in \mathbb{R}$. L'écart entre X et son projeté $\overline{X}$ sur Δ est égal à ce minimum. Or $||X - \overline{X}||^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \mathrm{Var}(X) = \sigma_X^2$.

L'écart-type (ou la variance) permet de mesurer l'écart entre une série statistique et sa moyenne et ainsi d'apprécier si elle peut être « bien » représentée par sa moyenne.

Autre résultat classique

La projection considérée étant orthogonale, on a d'après le théorème de Pythagore $||X||^2 = ||X - \overline{X}||^2 + ||\overline{X}||^2$

ce qui s'écrit aussi $\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + (\bar{x})^2$.

Par conséquent $\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2$ (formule de König).

2 - Coefficient de corrélation linéaire[4]

Soit un couple de séries statistiques (X , Y) dont les valeurs observées sur n individus d'une population sont représentées par deux vecteurs $X = (x_1, x_2, \ldots, x_n)$ et $Y = (y_1, y_2, \ldots, y_n)$ de $\mathbb{R}^n$.

---

[3] X est un vecteur de $\mathbb{R}^n$ or une illustration peut se faire (au mieux) dans $\mathbb{R}^3$. Cela peut donc fournir une aide à la compréhension dans un premier temps, tout en créant un obstacle pour la suite de l'apprentissage.

[4] On pourra mettre ce qui suit en parallèle avec l'analyse d'une série bivariée présentée dans l'article de Stéphan MANGANELLI, page 75.

Les projetés de X et Y sur la droite vectorielle Δ de vecteur directeur $\mathbb{1}$ sont les vecteurs $\overline{X}\,(\bar{x}, \bar{x}, \ldots, \bar{x})$ et $\overline{Y}\,(\bar{y}, \bar{y}, \ldots, \bar{y})$.

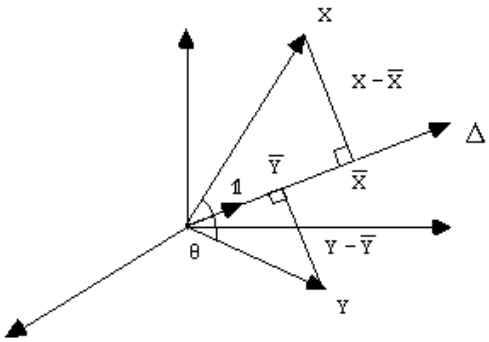


Figure 2

Si aucun des deux vecteurs X et Y n'appartient à Δ, les vecteurs $X - \overline{X}$ et $Y - \overline{Y}$ sont non nuls et orthogonaux à Δ ; ils forment un angle θ tel que :

$$\cos(\theta) = \frac{\langle X - \overline{X}, Y - \overline{Y} \rangle}{\|X - \overline{X}\| \, \|Y - \overline{Y}\|} = \frac{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

On reconnaît au numérateur l'expression de la covariance de X et Y, donc $\cos(\theta) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \rho$ où ρ est le coefficient de corrélation linéaire de la série statistique bivariée (X , Y).

- **Si $\cos(\theta) = \pm 1$:** alors les statistiques centrées $X' = X - \overline{X}$ et $Y' = Y - \overline{Y}$ sont des vecteurs colinéaires, c'est-à-dire qu'il existe un réel k tel que $Y - \overline{Y} = k\,(X - \overline{X})$. Alors, $Y = k\,X + \overline{Y} - k\,\overline{X}$ est de la forme $Y = k\,X + m\mathbb{1}$ (où $m \in \mathbb{R}$), ce qui signifie que X et Y sont en relation affine dans $\mathbb{R}^n$.
- **Si $\cos(\theta) \neq \pm 1$:** on est d'autant plus proche de la situation précédente que ρ est proche de 1 en valeur absolue.

III - Projection orthogonale sur un plan : régression linéaire simple

Si X et Y, supposés ne pas appartenir à Δ, sont liées par une relation du type $Y = aX + b\mathbb{1}$ ($a \in \mathbb{R}$ et $b \in \mathbb{R}$), notre problème de régression linéaire est résolu. Sinon, on cherche un vecteur $\hat{Y} = a\,X + b\mathbb{1}$ ($a \in \mathbb{R}$ et $b \in \mathbb{R}$) le plus proche possible de Y, c'est-à-dire tel que $\|Y - \hat{Y}\|^2$ soit minimum.

$\hat{Y}$ appartient à l'espace S engendré par X et le vecteur $\mathbb{1}$.

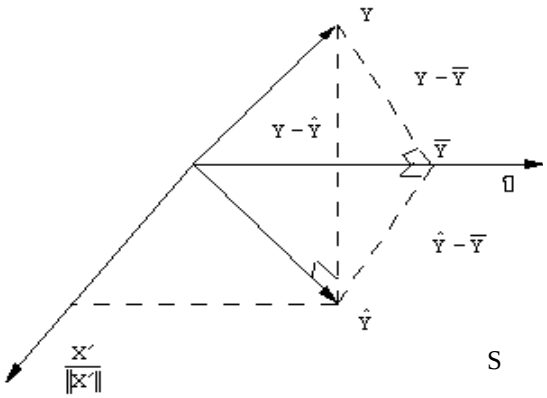


Figure 3

Le vecteur associé à la série statistique centrée $X' = X - \overline{X}$ appartient à S et est orthogonal à $\mathbb{1}$, il en est donc de même pour le vecteur $\frac{X'}{\|X'\|}$ qui de plus est de norme 1. Les vecteurs $\frac{X'}{\|X'\|}$ et $\mathbb{1}$ forment alors une base orthonormée de S.

Dans cette base, les coordonnées de $\hat{Y}$, projeté de Y sur S, sont $\langle Y , \frac{X'}{\|X'\|} \rangle$ et $\langle Y , \mathbb{1} \rangle = \bar{y}$.

On a ainsi $\hat{Y} = \langle Y , \frac{X'}{\|X'\|} \rangle \frac{X'}{\|X'\|} + \bar{y}.\mathbb{1} = \frac{\langle Y' + \overline{Y}, X' \rangle}{\|X'\|^2} X' + \overline{Y}$ où $Y' = Y - \overline{Y}$

$\hat{Y} = \frac{\langle Y' , X' \rangle}{\|X'\|^2} (X - \overline{X}) + \overline{Y}$ car $\overline{Y}$, qui appartient à Δ, est orthogonal à X'.

En remarquant que $\langle Y' , X' \rangle = Cov(X , Y)$ et $\|X'\|^2 = Var(X)$, on obtient :

$$\hat{Y} = \frac{Cov(X , Y)}{Var(X)} (X - \overline{X}) + \overline{Y}.$$

Ainsi $\hat{Y} = aX + b\mathbb{1}$ avec $a = \frac{Cov(X , Y)}{Var(X)}$ et $b = \bar{y} - a\,\bar{x}$.

Mesure de la qualité du modèle linéaire

$\hat{Y}$ étant le projeté orthogonal de Y sur S qui contient aussi $\overline{Y}$, $Y - \hat{Y}$ est orthogonal à $\hat{Y} - \overline{Y}$ (voir *figure3* pour une interprétation géométrique).

De l'égalité $Y - \overline{Y} = (Y - \hat{Y}) + (\hat{Y} - \overline{Y})$, on déduit $\|Y - \overline{Y}\|^2 = \|Y - \hat{Y}\|^2 + \|\hat{Y} - \overline{Y}\|^2$.

Comme la moyenne de $\hat{Y}$ est $\bar{y}$, $\|\hat{Y} - \overline{Y}\|^2 = \|\hat{Y} - \bar{y}\mathbb{1}\|^2 = Var(\hat{Y})$.

La quantité $\|Y - \hat{Y}\|$ est appelée *variance résiduelle* et l'égalité précédente devient :

$$\text{Var}(Y) = \text{Variance résiduelle} + \text{Var}(\hat{Y}).$$

Par définition, la solution $\hat{Y}$ est le vecteur affinement lié à X qui minimise la variance résiduelle, elle est donc de variance maximale.

$\hat{Y}$ est donc la variable appartenant à S, c'est-à-dire du type $aX + b\mathbb{1}$, qui explique la plus grande part de la variance de Y, elle rend maximum le rapport $\dfrac{\|\hat{Y} - \overline{Y}\|^2}{\|Y - \overline{Y}\|^2}$.

Or comme $(Y - \overline{Y}) - (\hat{Y} - \overline{Y}) = Y - \hat{Y}$ qui est orthogonal à $\hat{Y} - \overline{Y}$, le rapport précédent est égal à $\cos^2\left(\hat{Y} - \overline{Y}, Y - \overline{Y}\right)$, d'où :

$$\frac{\|\hat{Y} - \overline{Y}\|^2}{\|Y - \overline{Y}\|^2} = \frac{<\hat{Y} - \overline{Y}, Y - \overline{Y}>^2}{\|\hat{Y} - \overline{Y}\|^2 \|Y - \overline{Y}\|^2} = \frac{<a X + b - (a \overline{X} + b), Y - \overline{Y}>^2}{\| a X + b - (a \overline{X} + b)\|^2 \|Y - \overline{Y}\|^2} = \frac{<a (X - \overline{X}), Y - \overline{Y}>^2}{\|a (X - \overline{X})\|^2 \|Y - \overline{Y}\|^2}$$

$$= \frac{a^2 \text{Cov}^2(X, Y)}{a^2 \text{Var}(X) \text{Var}(Y)} = \rho^2.$$

Conclusion :

- **Si $\rho = \pm 1$:** $\|\hat{Y} - \overline{Y}\|^2 = \|Y - \overline{Y}\|^2$ donc $\|Y - \hat{Y}\|^2 = 0$ et $Y = \hat{Y}$, on est alors dans le cas d'une liaison affine entre X et Y.
- **Si $\rho = 0$:** $\|\hat{Y} - \overline{Y}\|^2 = 0$ donc $\hat{Y} = \bar{y}.\mathbb{1}$, par conséquent $\hat{Y}$ ne dépend pas de X. Le modèle linéaire n'explique rien de la variation de Y.

Dans tous les cas, la quantité ρ^2 mesure la part de variance de Y expliquée par le modèle linéaire.

IV - Projection sur un sous-espace : régression linéaire multiple

Soit une variable Y $(y_1, y_2, \ldots, y_n)$, appelée *variable expliquée*, et p variables $X_1, X_2, \ldots, X_p$, appelées *variables explicatives*[5] telles que $X_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,n})$.

On cherche le vecteur $\hat{Y}$ de l'espace S engendré par les vecteurs $X_1, X_2, \ldots, X_p$ et $\mathbb{1}$ le plus près possible de Y : $\hat{Y}$ est une combinaison linéaire des X_i et de $\mathbb{1}$ telle que $\|Y - \hat{Y}\|$ soit minimum.

---

[5] Cf. l'article de Stéphan MANGANELLI *Description d'une série statistique à deux variables quantitatives* page 75.

Écrivons $\hat{Y} = a_1X_1 + a_2X_2 + ... + a_pX_p + a_{p+1}1$. Cette relation peut être écrite sous forme matricielle :

$$\hat{Y} = Z\,A \text{ avec } Z = \begin{pmatrix} x_{1,1} & - & - & x_{p,1} & 1 \\ x_{1,2} & - & - & x_{p,2} & 1 \\ - & - & - & - & - \\ - & - & - & - & - \\ x_{1,n} & - & - & x_{p,n} & 1 \end{pmatrix} \text{ et } A = \begin{pmatrix} a_1 \\ a_2 \\ - \\ - \\ a_p \\ a_{p+1} \end{pmatrix}.$$

$\hat{Y}$ est le projeté de Y sur S. On peut démontrer (voir, par exemple **[2]** ou **[3]**) que les coefficients $a_1, a_2, ..., a_p$ et a_{p+1} sont donnés par $A = ({}^tZ\,M\,Z)^{-1}\,{}^tZ\,M\,Y$ où

$$M = \frac{1}{n} I_n \text{ et } Y = \begin{pmatrix} y_1 \\ y_2 \\ - \\ - \\ y_n \end{pmatrix}.$$

Comme en régression linéaire simple, on mesure la qualité du modèle à l'aide du rapport $R^2 = \dfrac{\|\hat{Y} - \overline{Y}\|^2}{\|Y - \overline{Y}\|^2} = \cos^2\left(\hat{Y} - \overline{Y}, Y - \overline{Y}\right)$.

Application avec p = 1 (régression linéaire simple)

Soient $X = \begin{pmatrix} x_1 \\ x_2 \\ - \\ - \\ x_n \end{pmatrix}$, $Y = \begin{pmatrix} y_1 \\ y_2 \\ - \\ - \\ y_n \end{pmatrix}$, $Z = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ - & - \\ - & - \\ x_n & 1 \end{pmatrix}$ et $A = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$.

On obtient ${}^tZ\,M\,Z = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & \frac{1}{n}\sum_{i=1}^n x_i \\ \frac{1}{n}\sum_{i=1}^n x_i & 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & \overline{x} \\ \overline{x} & 1 \end{pmatrix}$ et ${}^tZ\,M\,Y = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i\,y_i \\ \overline{y} \end{pmatrix}$

${}^tZ\,M\,Z$, de déterminant $\frac{1}{n}\sum_{i=1}^n x_i^2 - \left(\overline{x}\right)^2 = \mathrm{Var}(X)$, a pour inverse :

$$({}^tZ\,M\,Z)^{-1} = \frac{1}{\mathrm{Var}(X)} \begin{pmatrix} 1 & -\overline{x} \\ -\overline{x} & \frac{1}{n}\sum_{i=1}^n x_i^2 \end{pmatrix}.$$

On obtient alors $A = ({}^tZ\,M\,Z)^{-1}\,{}^tZ\,M\,Y = \dfrac{1}{\mathrm{Var}(X)} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i\,y_i - \overline{x}\,\overline{y} \\ -\overline{x}\,\frac{1}{n}\sum_{i=1}^n x_i\,y_i + \overline{y}\,\frac{1}{n}\sum_{i=1}^n x_i^2 \end{pmatrix}$,

donc $a_1 = \frac{1}{Var(X)}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\, y_i - \bar{x}\,\bar{y}\right) = \frac{Cov(X\,;\,Y)}{Var(X)}$

$$a_2 = \frac{-\bar{x}}{Var(X)}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\, y_i\right) + \frac{\bar{y}}{Var(X)}\,\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2}$$

$$= \frac{-\bar{x}\left[Cov(X\,,\,Y) + \bar{x}\,\bar{y}\right]}{Var(X)} + \frac{\bar{y}\left[Var(X) + (\bar{x})^2\right]}{Var(X)}$$

d'où $a_2 = -a_1\,\bar{x} - \frac{(\bar{x})^2\,\bar{y}}{Var(X)} + \bar{y} + \frac{\bar{y}\,(\bar{x})^2}{Var(X)} = \bar{y} - a_1\,\bar{x}$.

Conclusion

On retrouve le résultat du IV :

$$\hat{Y} = a_1 X + a_2 \mathbb{1} \quad \text{avec} \quad a_1 = \frac{Cov(X\,;\,Y)}{Var(X)} \quad \text{et} \quad a_2 = \bar{y} - a_2\,\bar{x}$$

V - Analyse en composantes principales[6]

On considère p variables quantitatives $V_1, V_2,..., V_p$ observées sur n individus $e_1, e_2,..., e_n$. Les p variables V_j peuvent être considérées comme p vecteurs de $\mathbb{R}^n$ tandis que les n individus e_i peuvent être considérés comme n points de $\mathbb{R}^p$.

Analysons les différences et les ressemblances entre les individus e_i $(1 \le i \le n)$ relativement à l'ensemble des caractères V_j $(1 \le j \le p)$. Pour cela, considérons les distances entre les points e_i. Dans la plupart des cas, ces distances dépendent des unités choisies pour exprimer les variables, on pallie cet inconvénient en travaillant avec les variables centrées réduites $X_1, X_2,..., X_p$ définies à partir des variables V_j par[7] :

$$X_j = \frac{V_j - \overline{V_j}}{\sigma(V_j)}.$$

On peut définir la matrice X suivante à partir des $n \times p$ mesures effectuées :

$$X = \begin{pmatrix} x_{1,1} & x_{2,1} & - & - & x_{p,1} \\ x_{1,2} & x_{2,2} & - & - & x_{p,2} \\ - & - & - & - & - \\ - & - & - & - & - \\ x_{1,n} & x_{2,n} & - & - & x_{p,n} \end{pmatrix}$$

dont les colonnes représentent les p caractères centrés et réduits, et dont les lignes représentent les n individus.

---

[6] Cf. l'article suivant qui présente le vocabulaire de la statistique exploratoire.

[7] On peut utiliser la démarche qui suit sur des données brutes (non centrées, réduites) mais certains des résultats numériques obtenus ne sont plus valables.

1 - Espace des individus

Il n'est pas facile d'apprécier la proximité de deux individus dans un espace de dimension $p \geq 3$. On cherche donc un sous-espace de dimension $q < p$ (en pratique $q = 1$ ou $q = 2$) sur lequel la projection des n individus sera la meilleure, autrement dit dans lequel les distances entre les individus seront les mieux conservées. La dispersion du nuage de points dans $\mathbb{R}^p$ est mesurée par la moyenne des carrés des distances entre individus que l'on appelle l'*inertie du nuage*. Comme les données sont centrées réduites, cette inertie est égale à p, le nombre de variables. Une projection réduisant toujours les distances, le critère pour trouver le meilleur sous-espace sera de maximiser les distances entre projetés.

On peut démontrer (voir, par exemple, **[2]** ou **[3]**) que la solution est le sous-espace dont la base orthonormée est formée de q vecteurs propres unitaires u_1, u_2, ..., u_q associés aux q plus grandes des p valeurs propres ($\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_p$) de la *matrice des corrélations* $R = {}^tX\,X$ où X est la matrice des données centrées réduites. Cette matrice R, carrée d'ordre p, est symétrique et on démontre que ses p valeurs propres sont toutes réelles positives et que leur somme est égale à la trace de la matrice R, donc à p. On montre également que l'inertie du nuage projeté sur l'axe engendré par le vecteur u_j est égale à λ_j ($1 \leq j \leq p$).

Ainsi, les solutions sont emboîtées, c'est-à-dire que le meilleur plan est engendré par les vecteurs propres associés aux deux plus grandes valeurs propres de la matrice des corrélations, c'est le *premier plan factoriel* ; le meilleur sous-espace de dimension 3 est engendré par les vecteurs propres associés aux trois plus grandes valeurs propres, etc. Il peut toutefois être intéressant de projeter sur les deuxième et troisième plans factoriels engendrés respectivement par u_1 et u_3 d'une part et u_2 et u_3 d'autre part.

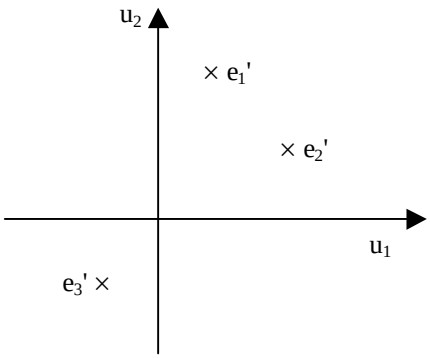
Le rapport

$$\frac{\Sigma\,(\text{distances entre les projetés})^2}{\Sigma\,(\text{distances entre les individus})^2} = \frac{\lambda_1 + \lambda_2 + ... + \lambda_q}{\lambda_1 + \lambda_2 + ... + \lambda_p} = \frac{\lambda_1 + \lambda_2 + ... + \lambda_q}{p}$$

représente le pourcentage d'inertie expliquée par le sous-espace. Il est une fonction croissante de q car les valeurs propres λ_j sont réelles positives.

Projeté d'un individu sur le premier plan factoriel

L'individu e_i se projette en e_i' dans le plan engendré par u_1 et u_2 de telle sorte que $e_i' = \langle e_i , u_1 \rangle u_1 + \langle e_i , u_2 \rangle u_2$.



Un individu est bien représenté dans ce plan si $\cos(e_i' , e_i) = \dfrac{\|e_i'\|^2}{\|e_i\|^2}$ est maximum c'est-à-dire si l'angle entre un individu et son projeté est minimum.

2 - Espace des caractères

À partir des coordonnées $(c_{i,1} , c_{i,2} , ... , c_{i,p})$ des e_i $(1 \le i \le n)$ dans la base des vecteurs propres u_i de la matrice R, on construit les variables synthétiques $C_1, C_2, ..., C_p$ où $C_j = (c_{1,j} , c_{2,j} , ... , c_{n,j})$. On montre que ces nouvelles variables, appelées *composantes principales*, sont les plus liées aux caractères de départ $V_1, V_2, ..., V_p$, qu'elles sont non corrélées entre elles, qu'elles sont de variance maximale et que ce sont des vecteurs propres de la matrice carrée d'ordre n, symétrique $X\,{}^tX$ (voir, par exemple, **[2]** ou **[3]**).

Les résultats sur les caractères sont liés à ceux obtenus sur les individus. En effet, les matrices ${}^tX\,X$ et $X\,{}^tX$ ont les mêmes valeurs propres non nulles et, u_j étant un vecteur propre unitaire de ${}^tX\,X$ associé à la valeur propre non nulle λ_j, $w_j = \dfrac{1}{\sqrt{\lambda_j}} X\,u_j$ est un vecteur propre unitaire de $X\,{}^tX$, associé à la même valeur propre λ_j, et colinéaire à C_j.

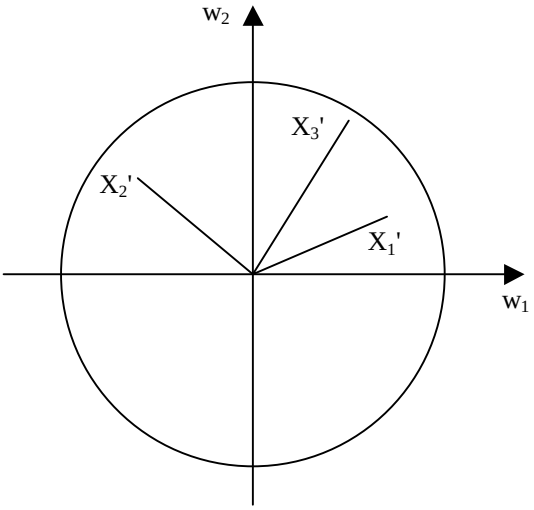
L'inertie des variables X_j par rapport à 0 est égale à p et l'inertie des projetés des variables X_j sur C_k est égale à λ_k. La qualité de la représentation des projetés des X_j sur l'espace de dimension q engendré par $C_1, ..., C_q$ $(q \le p)$ est égale à

$$\frac{\lambda_1 + \lambda_2 + ... + \lambda_q}{\lambda_1 + \lambda_2 + ... + \lambda_p} = \frac{\lambda_1 + \lambda_2 + ... + \lambda_q}{p}.$$

Projeté d'un caractère sur le premier plan factoriel

Le caractère X_i se projette en X_j' dans le plan engendré par les vecteurs propres w_1 et w_2. Le vecteur X_j' a pour coordonnées $<X_j , w_1>$ et $<X_j , w_2>$ dans la base (w_1 , w_2).

Toutes les variables X_j sont centrées réduites, donc ont pour variance 1. En travaillant avec la norme égale à la norme euclidienne divisée par n, introduite au début, les vecteurs X_j sont de norme 1, ils appartiennent à l'hypersphère de rayon 1. Les X_j' sont de norme inférieure à 1, il est judicieux de les représenter avec le cercle de rayon 1 appelé *cercle de corrélation*.



Qualité de la représentation

Plus le projeté X_j' est proche du cercle de corrélation, mieux la variable est représentée par son projeté.

La corrélation entre un caractère X_j et une composante principale w_k peut être appréciée sur le cercle de corrélation ou mesurée par $\cos(X_j , w_k) = \frac{<X_j , w_k>}{||X_j|| \, ||w_k||}$.

VI - Prolongements

L'analyse factorielle des correspondances (AFC) et l'analyse des correspondances multiples (ACM) qui traitent des tableaux de contingence, c'est-à-dire des mesures de p variables qualitatives sur n individus, sont des cas particuliers de l'analyse en composantes principales (ACP) ; l'analyse factorielle discriminante (AFD) qui a pour objet de répartir des individus entre les modalités

d'une variable qualitative (expliquée) suivant les mesures obtenues sur des variables quantitatives (explicatives) en est un autre. Elles sont donc également susceptibles de traductions géométriques (voir la présentation de ces méthodes dans l'article qui suit et de nombreux exemples dans **[4]**).

Certains concepts liés aux espaces euclidiens (norme, produit scalaire) ou à l'algèbre linéaire (changement de bases, valeurs et vecteurs propres) trouvent là une application qui permet de leur donner du sens. Réciproquement, la présentation dans ce cadre des concepts de statistique (moyenne, écart-type, régression, etc.), sera, pour certains, plus satisfaisant mathématiquement.

Ces considérations sont une justification, si besoin était, de la poursuite de l'étude de la géométrie dans les séries non scientifiques.

Bibliographie

[1] PIEDNOIR J.-L., DUTARTRE P., *Enseigner la statistique au lycée : des enjeux aux méthodes*, Brochure n° 112, de la commission Inter-IREM *Lycées Technologiques*, 2001.

[2] SAPORTA G., *Probabilités, Analyse des données et Statistique*, Technip, 1990.

[3] BOUROCHE J.-M., SAPORTA G., *L'analyse des données*, Que sais-je n° 1854, PUF, 1980.

[4] ROBERT C., *Analyse descriptive multivariée*, Flammarion médecine-sciences, 1989.

Différents domaines de l'analyse des données : techniques en statistique exploratoire[1]

Michel HENRY

L'étude systématique de vastes populations, groupes ou ensembles d'objets ayant des propriétés communes conduit au recueil de nombreuses données chiffrées, obtenues comme valeurs observées d'un ou plusieurs caractères appliqués à chaque élément (ou unité statistique) d'une telle population. Le but de la statistique exploratoire ou statistique descriptive, est de synthétiser, résumer, structurer l'information contenue dans les données. Elle utilise pour cela des représentations des données sous forme de tableaux, de graphiques, de schémas, d'indicateurs numériques.

L'exploitation massive de l'outil informatique a permis d'enrichir la statistique exploratoire de nombreuses techniques de visualisation de données multidimensionnelles. Ces techniques constituent les outils de l'analyse des données.

Analyse factorielle

En statistique descriptive, l'analyse factorielle regroupe l'ensemble des méthodes qui permettent d'interpréter des données issues de séries d'observations comme résultantes de l'effet de causes appelées facteurs que l'on se propose de séparer et mesurer. Conçue à l'origine pour l'analyse de tests psychométriques, l'analyse factorielle a été introduite par SPEARMAN (1904), KELLEY et THURSTONE. Depuis lors ces méthodes n'ont cessé de se développer et de se diversifier, notamment sous l'impulsion de HOTELLING (1933) en économétrie.

Analyse des données

L'analyse des données consiste à organiser, structurer et représenter des données issues de l'application d'un caractère multidimensionnel à une vaste population statistique, de manière à permettre une lecture simplifiée des informations principales relatives à cette population.

[1] Les présentations qui suivent des différentes techniques en statistique descriptive doivent beaucoup à l'ouvrage de Gilbert SAPORTA : *Probabilités, analyse des données et statistique.*

Les principales méthodes de l'analyse des données se répartissent en deux groupes :

- les méthodes de classification visant à réduire la taille de la population en formant des groupes homogènes ;
- les méthodes factorielles qui cherchent à réduire le nombre de variables, composantes du caractère étudié, en les résumant par un petit nombre de composantes synthétiques. Selon que l'on travaille avec des variables numériques ou qualitatives, on a recours à l'analyse en composantes principales ou à l'analyse des correspondances. Les liens entre deux groupes de variables peuvent être traités par l'analyse canonique.

Analyse en composantes principales

L'étude d'une population statistique de taille n passe le plus souvent par le recueil d'un nombre élevé p de données quantitatives par élément observé. L'analyse de ces données doit tenir compte de leur caractère multidimensionnel et révéler les liaisons existantes entre leurs composantes. L'analyse en composantes principales (ACP), introduite en 1901 par K. PEARSON et développée par H. HOTELLING en 1933, est une méthode très puissante pour explorer la structure de telles données. Chaque donnée étant représentée dans un espace à p dimensions, l'ensemble des données forme un « nuage de n points » dans $\mathbb{R}^p$. Le principe de l'ACP est d'obtenir une représentation approchée du nuage dans un sous-espace de dimension faible q par projection sur des axes bien choisis. Une métrique dans $\mathbb{R}^p$ étant choisie (en général normalisée par l'utilisation de variables centrées réduites), les q axes principaux sont ceux qui maximisent l'« inertie » du nuage projeté, c'est-à-dire la moyenne, éventuellement pondérée, des carrés des distances des points projetés à leur centre de gravité. Les composantes principales sont les p vecteurs ayant pour coordonnées celles des projections orthogonales des n éléments du nuage sur les q axes principaux. L'ACP construit ainsi de nouvelles variables, artificielles, et des représentations graphiques permettant de visualiser les relations entre variables, ainsi que l'existence éventuelle de groupes d'éléments et de groupes de variables. L'interprétation de ces représentations est délicate et doit respecter une démarche rigoureuse.

Analyse factorielle de dissimilarités

Les méthodes en analyse des dissimilarités ont le même objectif qu'en analyse en composantes principales (ACP) : trouver une configuration de n unités statistiques dans un espace de faible dimension, dans les cas où l'on ne connaît que les distances ou dissimilarités entre unités et non les variables les décrivant. Elles conduisent à des techniques originales. Dans le cas de distances euclidiennes dans un espace de dimension p inconnue, on peut calculer les composantes principales et faire une représentation euclidienne de l'ensemble des unités statistiques dans un

espace dont la dimension découle des données. Lorsque les dissimilarités entre les données ne sont pas des distances mais des mesures de proximité où l'information est de nature ordinale, on utilise des méthodes semi-métriques de positionnement qui consistent à rechercher une configuration des *n* points dans un espace euclidien de dimension fixée à l'avance (contrairement au cas précédent), telle que les distances entre ces points respectent au mieux l'ordre donné par les proximités pour le maximum de points.

Analyse canonique

Lorsque *n* individus sont décrits par deux ensembles de variables, on cherche à examiner les liens existant entre ces deux ensembles afin de savoir s'ils décrivent ou non les mêmes propriétés. Si ces deux ensembles sont confondus, un seul suffit pour la description statistique. Si ces deux espaces sont orthogonaux dans $\mathbb{R}^n$, c'est que les deux ensembles de variables appréhendent des phénomènes totalement différents. Entre ces deux cas extrêmes, on s'intéresse aux positions relatives de ces deux espaces de données en cherchant les éléments les plus proches. La démarche de l'analyse canonique qui consiste à rechercher des couples de variables en corrélation maximale est fondamentale, car elle se retrouve dans d'autres méthodes, notamment en analyse des correspondances ou en analyse discriminante. L'analyse canonique a été étendue à plusieurs ensembles de variables par P. HORST en 1961 puis J. D. CARROLL en 1968.

Analyse factorielle discriminante

Dans l'analyse en composantes principales (ACP), on a en vue la description d'un tableau de données de dimensions ($p \times n$) pour *p* caractères et *n* individus, les deux ensembles *I* des individus et *K* des caractères n'ayant aucune structure particulière. Dans l'analyse factorielle discriminante (AFD), développée par FISHER (1936), on se donne une partition sur l'ensemble des individus (on dit aussi, de façon équivalente, qu'on possède, outre les *p* caractères quantitatifs du tableau de données initial, un caractère qualitatif, avec un nombre fini *q* de modalités).

L'objet de l'analyse est alors de rechercher si ce caractère qualitatif supplémentaire possède une influence sur l'ensemble des *p* variables mesurées et de déterminer, le cas échéant, des caractères discriminants, c'est-à-dire des caractères induisant sur l'ensemble *I* des individus une partition aussi proche que possible de celle que définit la variable qualitative initiale. L'analyse factorielle discriminante se ramène à une analyse en composantes principales effectuée sur l'ensemble des centres de gravité des diverses classes d'individus correspondant aux *q* modalités de la variable qualitative.

Analyse factorielle des correspondances

L'analyse en composantes principales (ACP), et l'analyse factorielle discriminante (AFD) exploitent des tableaux de données contenant des caractères mesurés sur des individus. On est souvent en présence de tableaux différents, dont le contenu est formé par les fréquences avec lesquelles sont observées les modalités de deux caractères non nécessairement quantitatifs. Il s'agit de *tableaux de contingence* qui peuvent être d'assez grandes dimensions. En présence de tels tableaux, la statistique classique nous donne par le *test du Khi-deux* le moyen de savoir s'il existe une liaison (ou *correspondance*) entre les caractères étudiés, mais ne permet guère de décrire cette liaison, ce qui est précisément l'objet de l'analyse factorielle des correspondances (AFC) systématisée par J. P. BENZÉCRI en 1962. L'AFC peut être considérée comme une ACP particulière dotée de la métrique du Khi-deux qui ne dépend que du profil des colonnes du tableau. L'analyse permet, dans le plan des deux premiers axes factoriels, une représentation simultanée, souvent fort suggestive des ressemblances entre les colonnes ou les lignes du tableau et de la proximité entre lignes et colonnes.

Bibliographie très sommaire

APMEP, *Analyse des données*, publications de l'APMEP, tome 1, n° 28 et tome 2, n° 40.

BENZÉCRI, J.-P., *Histoire et préhistoire de l'analyse des données*, Cahiers de l'analyse des données, Dunod, 1976- I 1, 2, 3, 4, II 1.

BENZÉCRI, J.-P., *Analyse des données*, tome 1 : la taxinomie, tome 2 : Analyse des correspondances, Paris, Dunod, 1973.

DROESBEKE, J.-J., TASSI, P., *Histoire de la Statistique*, Paris, PUF, collection Que sais-je ? n° 2527.

LEBART, L., MORINEAU, A., PIRON, M., *Statistique exploratoire multidimensionnelle*, Paris, Dunod, 1997.

SAPORTA, G., *Probabilités, analyse des données et statistique*, Paris, Technip, 1990.

Deuxième partie :
simulations et modèles probabilistes

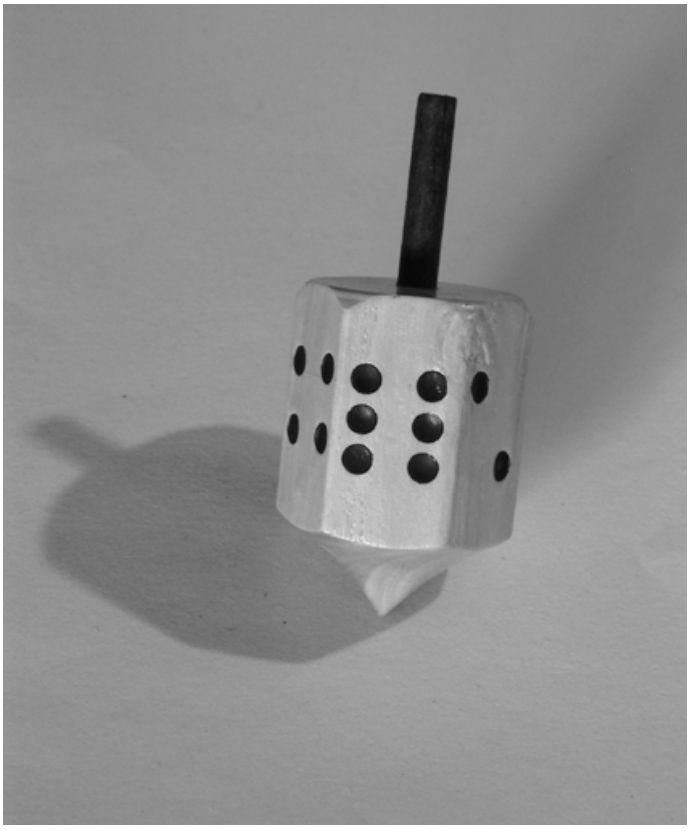
Dans cette deuxième partie, les questions posées par la simulation informatique font l'objet de notre travail d'approfondissement. En premier lieu, nous abordons successivement avec Jean-Claude GIRARD, Michel HENRY et Jean-François PICHARD les dimensions didactiques, épistémologiques et historiques du statut de la simulation.

Le fonctionnement du générateur aléatoire d'un ordinateur, pour mystérieux qu'il soit, est massivement mis à contribution dans les activités de simulation. Il ne va pas de soi et nécessite la compréhension en profondeur de ce qu'est une suite pseudo-aléatoire équirépartie, objet de recherches contemporaines en cours, présentées par Bernard PARZYSZ.

L'introduction de lois discrètes et continues dans le nouveau programme de terminale S et leur simulation nous a amené à proposer trois articles de Bernard PARZYSZ et Michel HENRY, présentant tour à tour les lois binomiale, exponentielle et normale.

Rappelant ensuite que toute courbe en cloche n'est pas nécessairement normale, Jean-François PICHARD étudie les liens historiques entre le théorème de Bernoulli, premier résultat en estimation, et l'aboutissement des travaux sur la théorie des erreurs, le Théorème-Limite Central, mettant en évidence la nature fondamentale des phénomènes gaussiens. Ce théorème confère à la loi normale un statut généraliste largement exploité en sciences sociales. Il est à la source des résultats de base dans les théories de l'estimation et des tests d'hypothèse.

La statistique inférentielle (techniques de décisions sur échantillons aléatoires), est une des branches majeures de la statistique, essentiellement développée au 20ème siècle avec les travaux de Karl PEARSON (test du Khi-deux, 1900) et les fondements théoriques de Jerzy NEYMAN et Egon PEARSON (1933). Notre Commission, par le biais d'universités d'été, a consacré plusieurs publications (référencées en bibliographie) à l'enseignement post-bac de ces deux théories principales. Les programmes de terminales introduisant le contrôle de l'adéquation de données statistiques expérimentales à une loi équirépartie, il nous a paru utile de présenter de manière très élémentaire une introduction aux tests d'hypothèses, suivie d'un article de Louis-Marie BONNEVAL et Michel HENRY donnant les éléments de théorie relatifs aux tests du Khi-deux, pour en dégager quelques remarques de nature didactique à l'intention des professeurs de terminale.



Modélisation et simulation en classe, quel statut didactique ?

Jean Claude GIRARD, Michel HENRY

I - Heurs et malheurs de la liaison statistique-probabilité dans l'enseignement secondaire

Les derniers changements dans les programmes de lycée (2000-2002) ont donné une importance beaucoup plus grande à l'enseignement de la statistique et ont introduit les probabilités comme outils théoriques de modélisation des situations aléatoires. La méthode pédagogique préconisée fait une large place à la simulation. Cette nouvelle approche pose de nouveaux problèmes didactiques liés aux difficultés propres à l'activité de modélisation et à l'ambiguïté du statut de la simulation. Elle ne permet pas de surmonter d'emblée les obstacles bien connus comme par exemple l'assimilation fréquence-probabilité.

La liaison entre l'enseignement de la statistique et celui des probabilités devrait être clarifiée. Elle nécessite sans doute de s'étendre sur de nombreuses années et pourrait s'inspirer de la progression de l'enseignement de la géométrie allant de l'école primaire à la terminale.

Si les probabilités ont été assez facilement considérées, dans l'enseignement comme ailleurs, comme une partie des mathématiques, il n'en a pas toujours été ainsi pour ce qui concerne la statistique ; prenons pour témoin le titre *Mathématiques et Statistique* d'un manuel de première[1] de 1966. Pour cette raison, et pour d'autres en rapport avec le mode de présentation préconisé par les programmes, la liaison statistique-probabilité a présenté différents visages, de réforme en réforme, jusqu'à la dernière en date.

L'enseignement des probabilités a été introduit de façon relativement récente dans le cursus secondaire (PARZYSZ, 1997) mais il a déjà connu bien des changements de cap. La réforme « des maths modernes », lui donnait en première et en terminale une forme axiomatique (espace probabilisé), la statistique étant alors considérée comme une application des concepts théoriques introduits abstraitement. A cette époque, et même après la réforme de 1981, la définition pratique de la probabilité était celle de LAPLACE, fondée sur l'hypothèse

[1] CLUZEL R., VISSIO P. et CHARTIER F., *Mathématiques et Statistique,* 1ère D, Delagrave, 1966.

d'équiprobabilité des issues élémentaires. Le calcul des probabilités se ramenait à des considérations plus ou moins sophistiquées de combinatoire. Certaines séries (C, par exemple) ne faisait d'ailleurs que cela.

A partir de 1986, les outils élémentaires de la statistique descriptive, systématiquement introduits en seconde le sont aussi progressivement dans toutes les classes du collège, mais en toute indépendance de la partie probabilités basée en terminale sur le *langage (logique) des événements* pour déboucher sur des calculs de dénombrements. En 1991, intervient la *révolution fréquentiste*. La notion de probabilité est introduite en première par l'observation de la stabilisation de la fréquence d'un événement lors de la répétition d'une expérience aléatoire. L'approche est expérimentale et la France lycéenne lance des punaises. On peut alors parler de probabilité non équirépartie. La combinatoire est renvoyée en terminale. Cette approche est donc clairement basée sur la liaison entre observation statistique et notion de probabilité. Mais la relation entre les concepts probabilistes théoriques et les outils pratiques de la statistique n'est pas poursuivie et le risque est alors que les élèves assimilent fréquence observée et probabilité.

La réforme des années 2000 est celle de la *révolution statistique*. D'abord en terme de volume horaire (1/8 de l'année, en seconde, dit le programme officiel) puis en raison des nouveautés présentées. Le programme de seconde introduit ainsi les fluctuations d'échantillonnage et la simulation. On n'y parle pas encore de probabilité, laissant s'exprimer la notion naïve de *chance*. Les élèves étudient les quartiles et les boîtes à moustaches en première et on y définit une loi de probabilité comme un objet théorique modélisant une distribution de fréquences observables lors de la répétition d'une même expérience aléatoire. En terminale, les lois de probabilités continues font leur apparition de même que le test d'adéquation d'une distribution observée à une loi équirépartie. Ces programmes montrent donc une volonté clairement affichée de lier statistique et probabilité mais deux difficultés didactiques apparaissent alors, dont l'origine se trouve dans les activités de modélisation et de simulation.

II - Les difficultés liées au processus de modélisation

L'hypothèse sous-jacente aux nouveaux programmes des lycées est que les élèves, confrontés en seconde aux fluctuations d'échantillonnage (réelles ou simulées, nous y reviendrons) seront *« aussi familiers avec les objets* distributions de fréquences *qu'ils le sont par exemple en sixième avec les objets* cubes *: après avoir effectivement manipulé des cubes en classes primaires, on peut leur parler de l'objet cube sans qu'ils en aient un devant les yeux : le cube est un objet mental familier*[2] *».*

[2]. Annexe 2 *A propos des probas-stat de première et terminale S*, projet de document d'accompagnement des nouveaux programmes de lycée, MEN 2001.

L'analogie est ainsi suggérée avec la géométrie. Mais suffit-il d'être familier avec les distributions de fréquences pour comprendre le concept de probabilité ? Si l'enseignement des probabilités présente une ressemblance avec celui de la géométrie (HENRY, 1999, GIRARD, 1999), on peut toutefois pointer une différence essentielle.

La construction du modèle euclidien en géométrie est un long processus (qui s'étend sur une dizaine d'années) et dont le point de départ est l'observation d'objets de la réalité. Ces objets sont reconnus globalement (à l'école maternelle et au début de l'école primaire), puis progressivement leurs propriétés sont repérées à l'aide d'instruments (à la fin de l'école primaire). Les objets mathématiques correspondants sont ensuite définis à partir de ces propriétés (au collège). On passe ainsi du carré reconnu de façon perceptive au concept de carré défini uniquement pas ses propriétés même si on en fait encore un dessin. On connaît la difficulté que ce saut conceptuel pose encore en quatrième malgré le nombre d'années sur lequel il s'étend, le passage de la réalité au modèle n'étant pas immédiat pour les élèves qui prennent progressivement (et douloureusement) contact avec la démarche scientifique.

Le programme prévoit pour la liaison statistique-probabilités une démarche analogue, mais celle-ci s'étend au mieux sur deux ans (puisqu'on ne parle pas de hasard ni d'expérience aléatoire au collège) et de façon assez tardive (15-16 ans) c'est-à-dire quand de nombreuses conceptions naïves se sont installées chez les élèves. Cet enseignement tardif n'arrange pas toujours les choses surtout quand il est de type dogmatique. En effet, la perception du hasard n'est pas univoque et met en jeu des croyances variées. L'observation de ses effets rencontre des obstacles, notamment en ce qui concerne les biais psychologiques bien étudiés, autrement plus subtils que pour la perception des objets géométriques comme le carré ou le cube, allant facilement au consensus pour leur interprétation en termes de configurations.

On peut pointer une autre différence. Le modèle euclidien étant construit, on peut se poser des problèmes dans le modèle (et on ne s'en prive pas) ou utiliser ce modèle pour traiter des problèmes concrets (ou faussement concrets) comme ceux bien classiques de la hauteur de la pyramide de Chéops ou du rayon de la terre. Ce type de question est toutefois rare en géométrie, en particulier en situation d'examen, sauf à expliquer en détail la modélisation qu'il faut en faire. On observe ainsi deux types de modélisation en géométrie. Le premier (construction d'un modèle) qui s'étend sur 10 ans et qui permet de passer de l'espace sensible à l'espace mathématique et un autre qui permet (ou permettrait) de résoudre des problèmes concrets, le modèle euclidien étant maîtrisé (en partie). La modélisation du deuxième type (utilisation d'un modèle), rare en géométrie, est par contre présente dans quasiment *tous* les exercices de probabilités. Ceux-ci sont en effet présentés avec un habillage concret. On fait l'hypothèse ici que les situations

proposées seront assez proches des situations d'apprentissage (c'est-à-dire de référence) pour que les élèves puissent, sans aide, effectuer le transfert d'abord d'un modèle équiréparti, et plus tard de la loi binomiale ou de certaines lois continues. Le risque majeur est alors que cette démarche n'ait pas beaucoup de sens pour beaucoup d'élèves, compte tenu du court temps d'assimilation.

La modélisation dans ce cas risque donc d'être fragile et peu étayée sauf à commencer plus tôt et s'étaler sur un temps plus long.

III - L'ordinateur, un instrument expérimental incontournable mais ambigu

Cette progression du perceptif au théorique est prise en charge par les documents d'accompagnement des programmes, de la seconde à la terminale. Le principe didactique est de multiplier les situations d'observations expérimentales en seconde, pour installer une familiarité avec ce qui sera pris comme base de l'édifice : les distributions de fréquences qui constitueront en classe de Première l'objet concret à théoriser sous la forme de loi de probabilité. Les moyens informatiques mis à la disposition des élèves leur permettent une approche expérimentale des fluctuations d'échantillonnage et du phénomène de *stabilisation* de la distribution des fréquences quand l'observation peut s'exercer sur un très grand nombre d'expériences.

Cet objectif expérimental ne peut être viable que grâce à l'exploitation massive des ordinateurs. En effet, si toute la classe s'y met, on peut à la rigueur récolter plusieurs centaines de résultats de lancers de dés réels par exemple, suffisamment pour mettre en évidence le phénomène de stabilisation de la distribution des fréquences autour de l'équirépartition. Mais, si l'on veut aller au-delà d'une simple approche qualitative, le contrôle de la valeur stabilisée f d'une fréquence d'une série d'expériences à l'autre avec une précision ε donnée, ainsi que celui du niveau de confiance accepté quand on donne une valeur p autour de laquelle ces fréquences observées doivent se répartir[3], passe par l'estimation expérimentale de la probabilité $P(|f-p|<\varepsilon)$, c'est-à-dire par la répétition un grand nombre de fois de cette série d'expériences, évidemment inaccessible manuellement. La puissance et la rapidité des générateurs de chiffres pseudo-aléatoires des outils informatiques le permettent aisément.

Mais l'ordinateur est-il un véritable générateur de hasard ? En principe, pas encore, tant que des phénomènes physiques, comme le bruit électronique par exemple, ne sont pas introduits à cette fin[4]. Actuellement, les nombres pseudo-

[3] Thème d'études suggéré en seconde

[4] En 1955, la Rand Corporation édita une table de nombres au hasard obtenue à partir de bruits de fond électroniques. Cf. « comment peut-on simuler le hasard ? » dans *Enseigner la statistique au lycée : des enjeux aux méthodes,* CII Lycées technologiques, p. 97.

aléatoires que l'ordinateur (ou la calculatrice) fournit sont déterminés, dès lors que le calcul a commencé à partir d'une initialisation. Mais la complexité de leur détermination, supposant un calcul lourd, rend impossible leur prévision par tout autre moyen humain. « *On a alors le phénomène fortuit* », selon l'expression de POINCARÉ. *Tout se passe donc comme si* les chiffres fournis par la fonction ALEA de l'ordinateur ou Random de la calculatrice étaient issus d'un tirage au hasard de boules numérotées de 0 à 9 dans une urne. La condition est que cette génération pseudo-aléatoire vérifie différents tests d'uniformité. Les moyens de contrôle à notre disposition ne permettent pas d'invalider cette hypothèse, et nous pouvons déclarer que l'ordinateur *simule* les tirages *au hasard* successifs de ces boules de l'urne. Le problème d'obtenir un tel comportement de l'ordinateur est un problème de spécialiste : on sait que la suite aléatoire de ces chiffres générés est en fait périodique sur une très longue période, mais pour notre usage, nous limitant à un nombre raisonnable de données, ce problème ne se pose pas.

Mais, souhaitant par exemple simuler un sondage où p est la proportion de réponses favorables à estimer, ayant introduit une valeur pour p dans l'ordinateur, avons-nous réellement simulé quelque chose ? Ou avons-nous seulement vérifié que la fréquence de réponses favorables dans un vaste échantillon est voisine de p[5], ce qui ne fait que confirmer que le fabricant de l'ordinateur et le concepteur du logiciel ont bien rempli leurs cahiers des charges ?

Cependant, ne négligeons pas l'intérêt de l'outil informatique. Dans notre exemple, il permet une approche expérimentale des situations de sondages, et si la proportion p est cachée aux élèves, nous avons un outil de résolution de problèmes jouant le même rôle que les calculettes graphiques quand elles tracent des courbes représentatives de fonctions données.

Au moyen d'un habillage éventuellement évocateur d'une situation aléatoire concrète, l'ordinateur est donc implicitement considéré comme un générateur de hasard qui, correctement programmé, fournit des résultats parfaitement représentatifs des issues que l'expérience concrète évoquée fournirait. Cette assimilation ne va pas de soi pour tous les élèves (cf. les biais psychologiques évoqués plus haut). L'ordinateur donne *au hasard* des chiffres de 1 à 6 équirépartis[6], mais *ce n'est pas un vrai dé*, même non pipé. L'équivalence pour être admise suppose implicitement l'idée que la probabilité de sortie de l'as sur le dé est égale à celle de sortie du 1 de l'ordinateur, c'est-à-dire que le concept de probabilité soit installé. L'utilisation de l'ordinateur pour introduire la notion de probabilité, comme représentant une expérience concrète, repose donc sur un cercle vicieux didactique, ou bien sur un contrat didactique préalable impossible à

[5] Mais également que la fréquence varie d'un échantillon à l'autre tout en étant, en général, différente de p.

[6] Passons ici sous silence la très grande difficulté de définir cette propriété (cf. l'article de Bernard PARZYSZ qui suit).

expliciter : c'est bien le hasard du dé qui est représenté par l'ordinateur. C'est à cette condition que la réalisation d'expériences aléatoires exploitant le générateur aléatoire de l'ordinateur peut être assimilée à une *simulation*, terme omniprésent dans les programmes. Il convient donc de cerner plus précisément le statut de la simulation informatique telle qu'elle est présentée en classe.

IV - Simulation et modélisation

La simulation est inséparable du concept de modèle[7]. Voici en effet la définition que donne l'Encyclopédie Universalis :

« La simulation est l'expérimentation sur un modèle. C'est une procédure de recherche scientifique qui consiste à réaliser une reproduction artificielle (modèle) du phénomène que l'on désire étudier, à observer le comportement de cette reproduction lorsque l'on fait varier expérimentalement les actions que l'on peut exercer sur celle-ci, et à en induire ce qui se passerait dans la réalité sous l'influence d'actions analogues ».

Une définition concernant directement notre objet est donnée par Yadolah DODGE dans l'entrée Statistique du Dictionnaire encyclopédique (Dunod, 1993)[8] :

« La simulation est la méthode statistique permettant la reconstitution fictive de l'évolution d'un phénomène. C'est une expérimentation qui suppose la constitution d'un modèle théorique présentant une similitude de propriétés ou de relations avec le phénomène faisant l'objet de l'étude ».

Rigoureusement, sans modèle théorique (puisque la notion de probabilité est exclue) il n'est pas question de simulation. Le terme semble abusif en seconde quand il s'agit seulement de *représenter* le comportement des issues d'une expérience concrète. Mais on peut penser qu'il y a un modèle sous-jacent, dès lors que la production des chiffres aléatoires par l'ordinateur n'est pas exploitée n'importe comment. Dans les situations simples, comme le lancer d'un dé ou tout autre générateur standard, l'équiprobabilité est implicitement admise et associée spontanément à la loi uniforme discrète censée gouverner la production des chiffres aléatoires. Le terme de *simulation* peut donc être justifié en référence à un savoir minimal probabiliste, mais pas explicitement à des élèves de seconde. Comment justifier à leurs yeux l'équivalence entre expérience aléatoire réelle ou pseudo-concrète (c'est-à-dire déjà idéalisée) et sa simulation informatique, judicieusement programmée à partir d'un modèle théorique ? Ce qui peut garantir cette équivalence, c'est que les deux expériences relèvent du même modèle probabiliste, indisponible à ce moment-là. Comment interpréter alors les fluctuations

[7] On trouvera une explicitation de ce concept et une étude didactique du processus de modélisation, dans le livre *Autour de la modélisation en probabilités*

[8] Cité dans la brochure *Simulation et statistique en seconde* de la Commission Inter-IREM *Lycées technologiques*, p. 10

d'échantillonnage observées dans la répétition de l'expérience simulée comme inhérentes au caractère aléatoire de l'expérience étudiée ? Sans réponses à ces questions, le professeur ne peut que s'en tenir à un écran de fumée supposant que les élèves ne se les poseront pas (ce qui n'est pas une hypothèse absurde).

Notre question est donc fondamentalement didactique. On ne peut se satisfaire de la seule exploitation de la puissance et de la rapidité de l'ordinateur permettant de présenter aux élèves une grande richesse de nouvelles expériences aléatoires, car cela n'aurait qu'un intérêt limité. Son intérêt didactique, en tant qu'outil de simulation, tient plus essentiellement en ce qu'il nous oblige à analyser la situation aléatoire en jeu, à émettre des hypothèses de modèle (par exemple sur le choix de la valeur de la probabilité de Bernoulli à implanter) et à traduire ces hypothèses en instructions informatiques, pour que l'ordinateur nous permette de résoudre des problèmes éventuellement inaccessibles par le calcul a priori. Cela suppose de comprendre le processus de modélisation et d'interpréter les résultats obtenus en les rapportant aux hypothèses de modèle introduites. Simulant ainsi une expérience aléatoire réelle, au-delà de la production de données statistiques, l'ordinateur utilisé dans ces conditions serait un outil didactique majeur pour l'apprentissage de la modélisation en probabilités.

Ce hiatus didactique est abordé de front dans le document d'accompagnement du programme de première[9] :

« On clarifiera brièvement les positions respectives de la modélisation et de la simulation : modéliser consiste à associer un modèle à des données expérimentales, alors que simuler consiste à produire des données à partir d'un modèle prédéfini. On parlera de simulation d'une loi de probabilité P ; la simulation d'une telle loi avec des listes de chiffres au hasard ne peut se faire que si P peut être construite comme loi image d'une loi équirépartie. Pour simuler une expérience on associe d'abord un modèle à l'expérience en cours, puis on simule la loi du modèle ; on pourra détailler ces étapes, sans cependant le faire systématiquement dans les cas simples des expériences de référence ».

Afin de préciser par un exemple ce qu'il entend par modèle, ce commentaire fort pertinent ne peut échapper à une référence à un concept théorique, celui de loi de probabilité en l'occurrence. Il souligne par là l'importance didactique du support théorique à la conceptualisation, et en fin de compte pour l'acquisition d'un véritable savoir de nature scientifique. Dans cet esprit, la modélisation est ainsi précisée :

« Modéliser une expérience aléatoire, c'est associer à cette expérience une loi de probabilité sur l'ensemble des issues possibles. Ce choix, c'est à dire la modélisation de l'expérience, est en général délicat à faire, sauf dans certains cas

[9] Accompagnement des programmes de lycée, Mathématiques, rentrée 2002, Ministère de la jeunesse, de l'éducation et de la recherche, CD-ROM.

où des considérations propres au protocole expérimental conduisent à proposer a priori un modèle. Il en est ainsi des lancers de pièces ou de dés pour lesquels des considérations de symétrie conduisent au choix d'un modèle où la loi de probabilité est équirépartie. On se restreindra donc aux expériences de référence en évitant tout discours général sur ce qu'est ou n'est pas la modélisation ».

Mais des garde-fous sont installés pour ne pas faire de la modélisation un objet de travail à ce niveau :

« En dehors de tels cas où des considérations quant à la nature des expériences permettent de proposer un modèle, le choix d'un modèle à partir de données expérimentales est beaucoup plus délicat et ne sera pas abordé dans l'enseignement secondaire. On se contentera, si nécessaire, de fournir un modèle en indiquant que des techniques statistiques ont permis de déterminer et de valider un tel modèle ».

Ainsi le lien avec les données statistiques n'est pas oublié, conformément à l'option expérimentaliste de ce programme. Il est à nouveau rappelé :

« La modélisation ne relève pas d'une logique du vrai et du faux : un modèle n'est ni vrai ni faux : il peut être validé ou rejeté au vu de données expérimentales. Une des premières fonctions de la statistique dite inférentielle est d'associer à une expérience aléatoire un modèle, ou une gamme de modèles compatibles en un certain sens à définir avec les données expérimentales dont on dispose, et aussi de définir des procédures de validation d'un modèle ».

Et le document d'accompagnement ajoute :

« Pour déterminer et/ou valider un modèle probabiliste, le premier outil dont on dispose est un théorème de mathématiques que l'on appelle loi des grands nombres*, dont un énoncé intuitif est : dans le monde théorique défini par une loi de probabilité P sur un espace E, les fréquences des éléments de E dans une suite de n expériences identiques et indépendantes tendent vers leurs probabilités quand n augmente indéfiniment ».*

Dans cette démarche consistant à donner du sens aux concepts probabilistes de base comme ceux de probabilité, de loi, de variable aléatoire, d'espérance mathématique et d'écart-type par l'observation statistique, la loi des grands nombres va effectivement jouer un rôle déterminant. Il convient de préciser là encore son statut.

V - Approche fréquentiste et loi des grands nombres

Le phénomène de stabilisation des fréquences est un fait d'observation. On le range parmi *les lois du hasard* au sens du terme de loi en physique. Il tient en fait beaucoup à la définition même de la fréquence : il est clair qu'au 1 000ème lancer d'une pièce, le pile obtenu aura beaucoup moins d'effet sur la fréquence des piles déjà obtenus qu'au 10ème lancer. Ce phénomène, repéré depuis l'antiquité, tout en

permettant aux joueurs d'évaluer les enjeux et organiser leurs mises[10], ne débouche pas automatiquement sur le concept de probabilité, seulement apparu dans la deuxième moitié du XVII[e] siècle. D'ailleurs celui d'espérance de gain l'a précédé chez PASCAL et HUYGENS. Pour FERMAT, en 1654, une face d'un dé à trois faces produit « *un tiers des hasards* » et Jacques BERNOULLI considère dans Ars Conjectandi que tous les événements naturels sont comme s'ils avaient été tirés d'une grande urne. MONTMORT en 1705 et DE MOIVRE en 1733 donneront de la probabilité la définition traditionnelle du rapport du nombre des cas favorables à celui de tous les cas possibles, supposés équiprobables.

Une définition en termes de fréquence stabilisée[11] rencontre de grandes difficultés épistémologiques, puisqu'elle prétendrait caractériser un objet mathématique (donc abstrait) à partir d'une donnée expérimentale, confondant par là le champ de la réalité avec le domaine des mathématiques. Un grand pédagogue des probabilités, Alfred RENYI, semble cependant adopter ce point de vue dans son manuel[12]. :

« Nous appellerons probabilité d'un événement le nombre autour duquel oscille la fréquence relative de l'événement considéré… Nous considérons donc la probabilité comme une valeur indépendante de l'observateur, qui indique approximativement avec quelle fréquence l'événement considéré se produira au cours d'une longue série d'épreuves… »

Ayant ensuite démontré le théorème dit de *la loi des grands nombres*, RENYI pointe l'adéquation de la théorie probabiliste :

« Ce fait, la stabilité de la fréquence relative, vient d'être démontré mathématiquement. Il est remarquable que la théorie rende possible une description précise de cette stabilité ; cela témoigne sans aucun doute en faveur de sa puissance »,

et il essaie de justifier ce qui apparaît comme un cercle vicieux :

« Nous avons en effet défini la probabilité grâce à la stabilité de la fréquence relative, mais d'autre part la notion de probabilité intervient pour caractériser cette stabilité. En réalité, il s'agit pourtant de deux choses entièrement différentes. La définition *de la probabilité comme valeur autour de laquelle oscille la fréquence relative n'est pas une définition mathématique mais une description du substrat concret du concept de probabilité. La loi des grands nombres de Bernoulli*

---

[10] Au XIII[ème] siècle, un poème, De Vetula, en tire des conseils pour parier sur la somme des points obtenue avec trois dés, problème que GALILÉE résoudra au début du XVII[ème] pour le Grand Duc de Toscane. Le livre de CARDAN écrit au XVI[ème] siècle, décrit la combinatoire des jeux de dés à cet effet.

[11] Une telle définition a été tentée par VON MISES en 1928, qui a bâti une axiomatique inspirée de la convergence des fréquences, sans grand succès face au modèle de KOLMOGOROV de 1933.

[12] Calcul des probabilités, Dunod, 1966.

par contre est fondée sur la définition mathématique de la probabilité et par conséquent il n'y a là aucun cercle vicieux ».

La *définition mathématique* évoquée par RENYI est celle d'une mesure sur un ensemble abstrait. La démonstration reste donc au sein du modèle mathématique et la loi des grands nombres en est un théorème. Son assimilation au phénomène de stabilisation des fréquences relèverait alors d'une confusion épistémologique.

Un énoncé rigoureux de la loi des grands nombres[13], même sous sa forme la plus simple du théorème de Bernoulli, suppose donc une *définition mathématique* de la probabilité et ne peut s'inscrire dans une confusion entre modèle et réalité. Cette définition peut être basée sur l'équiprobabilité pour un ensemble fini de cas, en supposant que pour toute expérience aléatoire, les issues observables sont constituées de tels cas. C'est une hypothèse qui peut paraître restrictive. Elle peut être aussi génératrice d'obstacles didactiques et il vaut mieux ne l'expliciter que dans les situations où cette hypothèse de modèle va de soi.

L'énoncé lui-même ne va pas sans difficultés de compréhension. Considérons la forme la plus simple de Bernoulli :

Lors de la répétition indéfinie d'une même épreuve de Bernoulli (i. e. à deux issues, succès de probabilité p ou échec de probabilité 1 − p) la probabilité que la fréquence des succès obtenus en n épreuves s'écarte de p de plus qu'un ε donné, tend vers 0 quand n tend vers l'infini.

Dans cet énoncé, deux probabilités conceptuellement bien différentes[14] apparaissent :

1 - la probabilité (objective) *p* d'un succès dans l'épreuve de Bernoulli, concevable comme la proportion de boules blanches dans une urne de Bernoulli ad hoc,

2 - la probabilité $P(|f-p|>\varepsilon)$, qui peut être considérée comme un contrôle (subjectif) des données expérimentales, reliant la fréquence *f* observée à la probabilité théorique p du succès.

Cette expression est source d'obstacle épistémologique car elle pourrait laisser entendre qu'il existe un énoncé intégrant dans une seule formule un objet de la réalité et un objet théorique. Mais la fréquence *f* dont il s'agit ici découle de

---

[13] La loi faible s'énonce ainsi :
Si $(X_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires définies sur Ω, indépendantes, d'espérances *m* et de variances finies telles que $\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i)\xrightarrow[n\to\infty]{}0$, alors la suite des $\overline{X_n}=\frac{1}{n}\sum_{i=1}^{n}X_i$ converge en probabilité vers *m*.

Remarques : Il suffit pour cela que les $\mathrm{Var}(X_i)$ soient bornées ou plus simplement toutes égales à σ^2.

On dit que la suite (Y_n) converge en probabilité vers *Y*, si pour tout $t>0$, $P(|Y_n-Y|>t)\xrightarrow[n\to\infty]{}0$.

[14] A ce point que LAPLACE utilisait deux mots : possibilité et probabilité.

l'expression d'une variable binomiale, définie sur l'espace probabilisé pour *représenter* le nombre de succès obtenu en *n* épreuves. Il s'agit donc bien d'un objet du modèle, et le théorème n'est qu'une conséquence des propriétés des coefficients binomiaux intervenant dans la loi de cette variable.

L'absence d'une définition officielle de la probabilité et cette difficulté épistémologique expliquent que le programme de première n'a pu intégrer un énoncé explicite de la loi des grands nombres. Comme ce théorème est la clé de la démarche de modélisation entreprise, puisque le document d'accompagnement lui confère le pouvoir de valider un modèle associé à une expérience aléatoire, réelle ou simulée, les concepteurs de ce programme ont dû faire le choix d'un « *énoncé vulgarisé* ». Peu opératoire sans l'expression explicite de cette probabilité $P(|f-p|>\varepsilon)$, il ne peut jouer qu'un rôle qualitatif favorisant l'intuition.

La loi des grands nombres est donc un théorème interne à la théorie des probabilités. Les fréquences des éléments de *E* obtenues dans une simulation informatique sont-elles dans le modèle ou dans la réalité ? Dans le premier cas, on ne valide rien du tout puisqu'on ne sort pas du modèle, dans le deuxième cas comment les premières peuvent-elles *tendre* vers les secondes ? Il y a donc là une ambiguïté qui fait que l'on ne sait plus si on raisonne dans le modèle ou dans la réalité, et s'il est vrai que « *l'esprit statistique naît lorsque l'on prend conscience de la fluctuation d'échantillonnage* », l'observation des fluctuations de la distribution de fréquences lors de plusieurs simulations d'une même expérience aléatoire conduit-elle naturellement à accepter l'idée d'une loi théorique fixe liée à cette expérience ? L'apprentissage des probabilités par la loi des grands nombres et la simulation, n'est donc pas un long fleuve tranquille. Il nécessite de toute façon un temps d'apprentissage beaucoup plus long que le temps d'enseignement qui lui est actuellement consacré.

VI - Conclusion

La formation d'images mentales à propos de l'aléatoire est plus délicate et demande encore plus de temps que dans le cas de la géométrie. Il faut donc proposer aux élèves des activités propres à créer ces images mentales bien avant le lycée. Des recherches récentes ont montré que des élèves de collège pouvaient utiliser la simulation pour construire des expériences équivalentes (BORDIER, 1991) et qu'ils pouvaient modéliser des situations par analogie avec des urnes de Bernoulli (COUTINHO, 2001). Mais pour l'instant, rien n'est prévu concernant l'aléatoire au collège (GIRARD et al., 2001).

Le nouveau programme de l'école primaire envisage par contre cette possibilité:

« Quelques exemples de phénomènes aléatoires peuvent être proposés dans la perspective de faire apparaître des régularités (par exemple, lancers d'une pièce ou d'un dé, lancers de deux dés dont on fait la somme) ».

Les prochains programmes de collège combleront-ils le trou entre l'école primaire et le lycée au sujet de l'aléatoire ? Si c'était le cas, la liaison statistique-probabilité serait alors présente tout au long du processus de modélisation, de l'école primaire à la seconde. Par analogie avec ce qui est proposé pour la géométrie, le point de départ pourrait être l'observation, la construction, la reproduction, la description et la représentation d'expériences aléatoires.

L'enseignement de la géométrie établit à partir d'objets concrets une construction intellectuelle qui n'utilise que les propriétés d'objets définis mathématiquement c'est-à-dire de concepts. Le passage se fait sur une très longue période d'abord par une reconnaissance globale et perceptive des objets concrets puis une reconnaissance instrumentée des propriétés sur les objets ou leur représentation dessinée avant de parvenir au raisonnement hypothético-déductif sur les propriétés elles-mêmes c'est-à-dire sur les objets mathématiques. Comme le mot *carré* fait partie du langage courant avant de faire partie du langage mathématique et d'en être un concept, il pourrait en être de même pour le mot *probabilité*. De toute façon, ce mot est utilisé dans le langage courant, alors autant lui donner un sens le plus précis possible.

Les techniques de statistique descriptive du collège (moyennes, pourcentages, graphiques en barres, circulaires, en boîtes, en tiges et feuilles…) trouveraient naturellement leur place pour communiquer et résumer les résultats de ces expérimentations. Les exercices de statistique au collège ne porteraient plus uniquement sur des populations mais aussi sur des échantillons tirés de ces populations. La liaison entre les deux points de vue se construirait sur plusieurs années. Elle pourrait être l'étude d'un caractère (qualitatif ou quantitatif) sur les individus d'un échantillon *tiré au hasard* dans une population. Ceci pourrait se faire en tirant des étiquettes correspondant aux individus dans un chapeau, ensuite avec une table de nombres au hasard et enfin avec la touche Random de la calculatrice. Ces expériences illustreraient d'une façon perceptive et progressive la variabilité des résultats et les fluctuations d'échantillonnage. La notion d'expériences équivalentes se construirait alors en acte et le passage au modèle sous-jacent aurait plus de chance de prendre sens plus tard, si toutefois ce passage est clairement identifié comme objet d'enseignement.

Pour le moment, toutes ces étapes sont concentrées en seconde (souvent en fin d'année) et en première (éventuellement avant la définition formelle). Il y a peu d'études sur les difficultés des élèves par rapport à cette nouvelle présentation, mais on en a pu en avoir un aperçu lors des quelques formations qui ont été mises en place à la suite du changement de programme dans les difficultés des professeurs qui n'ont pas reçu d'enseignement sur le sujet pendant leurs études.

Bibliographie

BORDIER, J., *Un modèle didactique utilisant la simulation sur ordinateur, pour l'enseignement de la probabilité*, Thèse de doctorat, Université Paris-7, 1991.

COUTINHO, C., *Introduction aux situations aléatoires dès le collège : de la modélisation à la simulation d'expériences de Bernoulli dans l'environnement informatique Cabri-Géomètre 2*, Thèse de doctorat, Université Joseph Fourier, Grenoble, 2001.

C.R.E.M. : *L'enseignement des sciences mathématiques*, chap. 2, Statistique et Probabilités, p. 51-86, J.-P. KAHANE (coord.), Odile Jacob, Paris, 2002.

DUTARTE, P., KERN, C. coord., *Simulation d'expériences aléatoires*, Commission Inter-IREM Lycées Technologiques, IREM de Paris-Nord, 1998.

GIRARD, J. C. : Le professeur de mathématiques doit-il enseigner la modélisation ?, *Repères-IREM n° 36*, p. 7-14, Topiques Editions, 1999.

GIRARD, J. C., HENRY M., PARSYSZ B., PICHARD J.-F. : Quelle place pour l'aléatoire au collège, *Repères-IREM n° 42*, p. 27-43, Topiques Editions, 2001.

G.E.P.S. : *Document d'accompagnement des programmes, Mathématiques classes de première des séries générales*. C. ROBERT (coord.) Direction de l'enseignement scolaire, MEN, Paris, CNDP, 2001.

HENRY, M. : L'introduction des probabilités au lycée : un processus de modélisation comparable à celui de la géométrie, *Repères-IREM n° 36*, p. 15-34, Topiques Editions, 1999.

HENRY, M. coord, *Autour de la modélisation en probabilités*, CII Statistique et Probabilités, Besançon, PUFC, 2001.

PARSYSZ, B., L'enseignement de la statistique et des probabilités dans l'enseignement secondaire, d'hier à aujourd'hui, in *Enseigner les probabilités au lycée*, p. 17-38. Commission Inter-IREM Statistique et Probabilités, IREM de Reims, 1997.

PARSYSZ, B. : L'enseignement de la Statistique et des probabilités en France : évolution au cours d'une carrière d'enseignant (période 1965-2002), in *Probabilités au lycée*, Commission Inter-IREM Statistique et Probabilités, brochure APMEP n° 143, p. 9-34, 2003.

PIEDNOIR, J.-L., DUTARTE, P., *Enseigner la statistique au lycée : des enjeux aux méthodes*, Commission Inter-IREM Lycées Technologiques, IREM de Paris-Nord, 2001.

VERLANT, B. (coord.), *Simulation et statistique en seconde*, Commission Inter-IREM Lycées Technologiques, IREM de Paris-Nord, 2000.


18
14
2
20

Expérimentation et simulation probabiliste

Jean-François PICHARD

I - Mathématiques expérimentales, un exemple historique en calcul des probabilités

L'expression *mathématiques expérimentales* est utilisée ici en analogie avec ce qui se passe en physique où l'expérimentaliste va inférer des conclusions à partir de ses observations pour construire son modèle théorique du phénomène. Un des premiers exemples de mathématiques expérimentales est donné par ARCHIMÈDE dans *La quadrature de la parabole*[1] qui, après avoir obtenu une relation expérimentale entre un segment de parabole et un triangle associé par pesée de plaques minces ayant ces mêmes formes, fait une analyse théorique mécanique par application de la statique du levier, décrite dans son traité *De l'équilibre des figures planes* (*op. cit.*), puis une démonstration géométrique rigoureuse de cette relation à l'aide de la méthode d'exhaustion de EUDOXE et des propriétés des coniques. Cette approche expérimentale, que ARCHIMÈDE avait menée à son terme dans le cas de la quadrature de la parabole, se trouve aussi dans des questions liées au calcul des probabilités.

Le premier auteur sur le calcul des probabilités, Jérôme CARDAN[2], avait une longue expérience et une pratique assidue des jeux de hasard, mais il ne semble pas en avoir tiré une méthode pour estimer la valeur des chances. GALILÉE[3] fait une allusion à une indication donnée par l'expérience, concernant le jeu avec trois dés :

« *Toutefois, quoique le 9 et le 12 se composent en autant de façons que le 10 et le 11, ce pour quoi ils devraient être présumés d'usage égal, on voit néanmoins que la longue observation a fait estimer par les joueurs que le 10 et le 11 sont plus avantageux que le 9 et le 12.* »[4],

---

[1] ARCHIMÈDE (II^e^ siècle av. J.-C.). *Oeuvres*, tome II, traduit par C. MUGLER, éd. Les Belles Lettres, Paris, 1971.

[2] CARDANO, Gerolamo (vers 1560), *Liber de Ludo Aleae*, publié dans *Opera*, Lyon, 1663 ; trad. *The Book on Games of Chances*, Holt, Rinehart and Winston. New-York, 1961.

[3] GALILEI, Galileo (vers 1620). Considerazione sopra il Giuco dei Dadi, *Opere de Galileo Galilei*, Firenze, 1855, t. xiv, p. 293-296 ; 1^ère^ publication des *Opera*, Florence, 1718.

[4] Un poème du XIII^e^ siècle, De Vetula, décrit les 216 *manières de tomber* équipossibles, correspondant aux 56 combinaisons observables sur 3 dés réalisant les 16 totaux possibles. Il établit que les sommes 10 et 11 peuvent être obtenues de 27 manières, et l'emportent sur les 9 et 12, réalisés seulement par 25 manières de tomber.

mais le premier exemple d'une véritable expérimentation est celle faite par BUFFON en relation avec le problème de St Pétersbourg, que nous allons exposer brièvement.

Au début du XVIII^e^ siècle[5], après la parution de la première édition de son ouvrage *Essay d'analyse sur les Jeux de hazard*, MONTMORT[6] a entamé une correspondance avec Nicolas BERNOULLI[7] qu'il a publiée dans la deuxième édition de son livre (1713). Dans sa dernière lettre à MONTMORT du 9 septembre 1713 (pp.401-402, la seconde édition de l'*Essay* a été publiée en décembre 1713), Nicolas BERNOULLI proposait un certain nombre de problèmes sur des jeux de hasard. Citons un extrait (p.402) :

« Quatrième Problème :

A promet de donner un écu à B, si avec un dé ordinaire il amène au premier coup six points, deux écus s'il amène le six au second, trois écus s'il amène ce point au troisième coup, quatre écus s'il l'amène au quatrième, & ainsi de suite ; on demande quelle est l'espérance de B.

Cinquième Problème :

On demande la même chose si A promet à B de lui donner des écus en cette progression 1, 2, 4, 8, 16, &c. ou 1, 3, 9, 27, &c. ou 1, 4, 9, 16, 25, &c. ou 1, 8, 27, 64, &c. au lieu de 1, 2, 3, 4, 5, &c. comme auparavant. Quoique ces Problèmes pour la plûpart ne soient pas difficiles, vous y trouverés pourtant quelque chose de fort curieux. »

Dans sa réponse, MONTMORT indique (p. 407) :

« Les deux derniers de vos cinq Problêmes n'ont aucune difficulté, il ne s'agit que de trouver les sommes des suites dont les numérateurs étant en progression des quarrés, cubes, &c. les dénominateurs soient en progression géométrique : feu M. votre Oncle a donné la methode de trouver la somme de ces suites. »

Dans le cas où un jeu a un ensemble fini d'éventualités, HUYGENS[8] avait pris comme définition d'un jeu équitable un jeu dans lequel, quand les chances sont égales, la mise d'un joueur est égale à la « *valeur de sa chance* » (i.e., l'espérance mathématique de son gain). Pour un jeu associé à un temps d'attente

---

[5] Une étude plus détaillée des auteurs et des idées sur le calcul des probabilités est faite dans PICHARD J. F. : Les probabilités au tournant du XVIII^e^ siècle, dans *Enseigner les probabilités au lycée*, Commission Inter-IREM Statistique et Probabilités, éditeur IREM de Reims, 1997 ; reproduit dans *Autour de la modélisation en probabilités*, coord. M. HENRY, Presses Universitaires Franc-Comtoises, Collection Didactiques, 2001.

[6] MONTMORT, Pierre Rémond (1708). *Essay d'analyse sur les jeux de hazard* ; 2^e^ édition, 1713.

[7] Nicolas BERNOULLI était un neveu de Jakob BERNOULLI, il avait fait une thèse en droit à Bâle (Suisse) en 1709 utilisant des résultats de calcul des probabilités de son oncle Jakob et a participé à l'édition en 1713 de *Ars conjectandi* de Jakob BERNOULLI.

[8] HUYGENS, Christiaan (1657). *De ratiociniis in Ludo aleae* ; trad. Du calcul dans les jeux de hasard, in tome 14, *Oeuvres complètes*, 22 vol., 1888-1950, La Haye.

(éventuellement infini) d'un succès, qui a une loi géométrique de probabilité 1/6 (cas d'un dé parfait), la notion et la définition d'espérance mathématique sont étendues par MONTMORT et N. BERNOULLI directement du cas fini au cas infini. Nous pouvons noter aussi que la notion d'espérance n'était pas clairement dissociée de la définition de la probabilité.

Nous retrouvons un problème de ce même genre quelques années plus tard, mais la simplification d'utiliser une pièce au lieu d'un dé va rendre plus apparent un paradoxe singulier. Dans une lettre de 1728 à N. BERNOULLI, CRAMER[9], un mathématicien de Genève, mentionne le fait que le calcul donne une espérance de gain infinie pour le jeu avec une pièce et les gains en progression géométrique 1, 2, 4, 8, 16, etc., et donc que la mise de *B* devrait aussi être infinie pour que le jeu soit équitable. Sentant intuitivement que l'espérance, et donc la mise, est finie, CRAMER imagine pour surmonter cette difficulté des considérations morales pour faire « *accorder le calcul mathématique avec le bon sens* » :

(1) il considère que les valeurs de toutes les sommes d'argent plus grandes que 2^{24} sont pratiquement égales ; il trouve alors une somme de 13 écus pour la mise équitable ;

(2) il considère que le plaisir dérivé du gain d'une somme d'argent doit être pris comme variant avec la racine carrée de la somme et arrive alors à une mise équitable de $1/(2-\sqrt{2})^2$ associée à ce qu'il appelle espérance morale de gain de *B*.

Daniel BERNOULLI, dans un mémoire dans les Commentaires de l'Académie de Saint-Pétersbourg publié en 1738[10] (d'où le nom de *problème de St Pétersbourg*), proposait que l'avantage d'un gain soit proportionnel à la fortune initiale *a* du joueur, d'où la première intervention d'une équation différentielle $y = \frac{Cdx}{x}$ en calcul des probabilités, et il obtient une mise valant quelques écus et dépendant de la fortune *a* du joueur. BUFFON dans son *Essai d'arithmétique morale*[11] (1777), présente une histoire abrégée du problème de St Pétersbourg. Il constate que les méthodes proposées sont insatisfaisantes, par exemple dans le cas de l'espérance morale calculée selon la méthode de D. BERNOULLI, la mise devrait dépendre de la fortune du joueur, ce qui va à l'encontre de la notion de *justice* selon B. PASCAL et

---

[9] Gabriel CRAMER a donné son nom à un type de systèmes d'équations linéaires qu'il a étudié dans son ouvrage *Introduction à l'analyse des lignes courbes algébriques* (Genève, 1750).

[10] BERNOULLI, Daniel (1738). Specimen theoriae novae de mensura sortis. *Commentarii academiae scientarum imperialis Petropolitanae*, t.5 pour 1730-31. Traduit dans : Exposition of a new theory on the measurement of risk. *Econometrica*, Vol. 22 No.1, 1954. Daniel était le fils de Jean, dernier frère de Jakob et mathématicien connu ; Jean a été professeur à Groningen (Hollande), puis à Bâle.

[11] BUFFON, G.-L. LECLERC de (1777). *Essai d'arithmétique morale*, dans *Oeuvres Complètes*, tome 12, Ed. Garnier, 1855, reproduit avec des commentaires dans BINET J.-L. et ROGER J., *Un autre Buffon*, Paris, Hermann, 1977.

de la définition d'un jeu équitable donnée par HUYGENS. Les solutions qui, soit posent une valeur maximum à la somme versée ou comptent pour nulle une probabilité très petite, sont aussi mal assurées et conduisent à des valeurs différentes de la mise pour un jeu équitable. Il propose alors de faire une expérimentation pour avoir une valeur approximative de cette mise équitable, se basant sur une extension du théorème de Jakob BERNOULLI (§18) :

« le premier moyen qui se présente est de comparer le calcul mathématique avec l'expérience ; car dans bien des cas, nous pouvons par des expériences réitérées, arriver à connaître l'effet du hasard, aussi sûrement que si nous le déduisions immédiatement des causes[12] *... J'ai donc fait 2048 expériences sur cette question, c'est-à-dire, j'ai joué 2048 fois ce jeu en faisant jeter la pièce en l'air par un enfant ; les 2048 parties de jeu ont produit 10.057 écus en tout, ainsi la somme équivalente à l'espérance de celui qui ne peut que gagner, est à peu près cinq écus pour chaque partie. »*

Nous voyons ici la première application en calcul des probabilités d'une méthode expérimentale pour déterminer de façon approchée une valeur qui ne peut pas être obtenue directement par calcul avec la théorie connue. Tous les grands auteurs sur le calcul des probabilités du milieu du XVIII^{e} au début du XX^{e} siècle ont traité ce problème de St Pétersbourg, essentiellement avec les mêmes arguments cités :

- limitation du nombre de jeux d'une partie associé au nombre maximal du gain à payer,
- compter pour négligeable, et par suite nulle, une probabilité très petite,
- utilisation de la théorie de l'espérance morale de Daniel BERNOULLI.

Citons quelques auteurs qui ont traité ce problème dans les deux siècles suivants, CONDORCET[13], LAPLACE[14] et BOREL[15]. En particulier, BOREL a remarqué que si, pour une partie, le gain X est une variable aléatoire ayant une espérance mathématique μ et une variance finies, le jeu est *équitable* au sens classique du terme si la mise est égale à μ, pourtant, au bout de *n* parties identiques et

---

[12] Ce que propose BUFFON est de faire une application du théorème de Jakob Bernoulli, appelé maintenant *loi des grands nombres*. Voir BERNOULLI, Jakob (1713). *Ars Conjectandi, 4ème partie*, traduit du latin par Norbert MEUSNIER, IREM de Rouen, 1987.

[13] CONDORCET, Jean Nicolas de CARITAT, Marquis de (1784). Mémoire sur le calcul des probabilités, *Mémoires de mathématiques et de physique de l'Académie Royale des Sciences*, Paris. Voir aussi pp.392-397 dans Condorcet, *Arithmétique politique, Textes rares ou inédits (1767-1789)*, Ed. B. BRU et P. CREPEL, INED, 1994.

[14] LAPLACE, Pierre Simon de (1812). *Théorie analytique des probabilités*, chapitre X. De l'espérance morale, in *Oeuvres Complètes*, Tome 7, Paris, 1886 ; voir aussi *Essai philosophique sur les probabilités*, (5^{e} édition de 1825), Bourgois, 1986.

[15] BOREL, Emile (1938), *Valeur pratique et philosophie des Probabilités*. 2ème édition, 1952. Voir aussi diverses notes aux *Comptes Rendus* dans *Œuvres de Borel*, 4 vol., CNRS, Paris, 1972.

indépendantes, le gain (ou la perte) total net $\sum_{k=1}^{n}(X_k - n\mu)$ peut devenir très grand, de l'ordre de $\sqrt{n}$. Mais ce n'est pas avant 1937 que le problème de St Pétersbourg a reçu une solution compatible avec le concept d'espérance mathématique : W. FELLER[16] a élaboré une nouvelle définition d'un jeu équitable, qui généralise la notion établie par HUYGENS. FELLER prend une mise m_n qui dépend du nombre n de parties et appelle *équitable* un jeu pour lequel, en posant $S_n = \sum_{k=1}^{n} X_k$, on a : $\forall\ \varepsilon > 0,\ P\left(\left|\frac{S_n}{m_n} - 1\right| > \varepsilon\right) \to 0$. Dans le cas où le gain X_k d'une partie a une espérance mathématique finie μ et où il y a une mise constante égale à μ, alors $m_n = n\mu$ et la condition indiquée est vraie en raison de la loi des grands nombres. Pour le jeu de St Pétersbourg, où le gain X est une variable aléatoire de loi $P(X = 2^k) = 1/2^k,\ k \in \mathbf{N}$, donc ayant une espérance mathématique infinie, Feller démontre (en utilisant la méthode de troncature de la théorie de l'intégration) que le jeu est équitable pour $m_n = n\mathrm{Log}_2 n$. Ici la mise pour une partie individuelle est $\mathrm{Log}_2 n$, dépendant du nombre n de parties à jouer, chacune de ces parties individuelles pouvant durer infiniment longtemps. Cette solution ne répond donc pas tout à fait à la question que se posaient les savants du XVIII[e] siècle, à savoir déterminer la mise équitable pour une partie.

Quelques autres expérimentations concernant une épreuve aléatoire

Indiquons d'abord le problème de l'aiguille, dit de Buffon[17]. Le jeu consiste à lancer *au hasard* une aiguille de longueur a sur un parquet formé de lames équidistantes de largeur b et à observer si l'aiguille coupe ou non une rainure du parquet. BUFFON a montré, en utilisant le calcul différentiel et intégral inventé alors depuis peu de temps, que la probabilité que l'aiguille coupe une rainure est $2a/\pi b$. L'expérimentation sur ce jeu de l'aiguille de Buffon a été à la mode dans la seconde moitié du 19[e] siècle pour obtenir une approximation de π (c'est un exemple de la méthode de Monte-Carlo, voir ci-après). Mais ce n'était qu'un passe-temps, au mieux une vérification de la loi des grands nombres, car une valeur de π plus précise que celle obtenue par cette méthode était connue depuis longtemps[18].

---

[16] FELLER, William (1950). *An Introduction to Probability Theory and Its Applications*, Vol. I, 3ème édition, John Wiley & Sons, New York, 1968, chap. X, n°4.

[17] BUFFON avait étudié cette question dans un mémoire non publié de 1733, ce qui lui valut d'être admis comme assistant mécanicien à l'Académie Royale des Sciences, voir son *Essai d'arithmétique morale* (1777).

[18] Voir ARCHIMEDE, *La mesure du cercle*, *op. cit.* Voir aussi, entre autres, *Le nombre* π, ADCS, 1992 ; et DELAHAYE, Jean-Paul (1997). *Le fascinant nombre* π, Pour la Science, Belin, Paris.

Citons aussi une expérimentation qui était d'abord à but pédagogique. Pour une présentation à une conférence qu'il a donnée en 1874, Francis GALTON a imaginé un dispositif physique pour illustrer le principe de la loi des erreurs ou de dispersion, appareil qu'il nommait *quincunx*[19] et appelé plus tard *planchette de Galton*[20]. Des chevilles sont fixées perpendiculairement en quinconce sur des rangées régulièrement espacées sur une planchette et, cette planchette étant posée verticalement, on fait tomber des petites billes de plomb d'un entonnoir placé en haut de l'appareil ; ces billes, coulant entre une vitre et la planchette du fond, rebondissent sur les chevilles des rangées successives à gauche ou à droite et en arrivant dans les compartiments au bas de l'appareil dessinent la forme d'une distribution qui est binomiale en première approximation. L'appareil fabriqué pour GALTON comportait 19 rangées de chevilles placées de telle sorte que les probabilités qu'une bille rebondisse à gauche ou à droite étaient presque égales et la distribution obtenue dans les compartiments au bas de l'appareil était presque une courbe normale. Par la suite, avec un appareil ayant un nombre plus grand de rangées de chevilles, GALTON a introduit un niveau intermédiaire à compartiments ; les distributions (conditionnelles) dans ceux-ci étaient presque normales et en ouvrant chacun des compartiments la distribution au bas de l'appareil était encore presque normale. Ce dispositif a permis à GALTON de concevoir une distribution normale comme obtenue par superposition (mélange) de distributions normales, résultat qu'il a utilisé dans ses travaux sur l'hérédité. Peu après, Karl PEARSON[21] a modifié le dispositif en changeant l'orientation des chevilles pour obtenir une distribution binomiale dissymétrique.

A l'encontre de ce qu'a fait BUFFON, qui s'est contenté de la valeur numérique donné par son expérimentation (mais le problème était difficile), l'appareil inventé par GALTON lui a permis d'apercevoir une modélisation du mélange d'influences sur des caractéristiques humaines par hérédité, qu'il a ensuite exploitée, et que Karl PEARSON a développée en biométrie.

II - La simulation

1 - Quelques définitions du terme *simulation*

La partie *statistique et probabilités* du programme de mathématiques au lycée fait une large place à la simulation, en particulier pour l'introduction en seconde de la notion de variabilité et de fluctuation. La simulation est, à ma connaissance, très

[19] Cet appareil est décrit dans GALTON, Francis (1889). *Natural Inheritance*, Macmillan, London ; voir aussi STIGLER, Stephen M. (1986). *The History of Statistics. The Measurment of Uncertainty before 1900*, Bellknap Harvard.

[20] Cf. l'article de Bernard PARZYSZ : Du modèle à sa réalisation. La planche de Galton réalise-t-elle vraiment une distribution binomiale ?

[21] Voir PEARSON, Karl (1914-1930). *The Life, Letters and Labours of Francis Galton* (3 vols), Cambridge, Cambridge University Press.

peu abordée dans le cursus universitaire des enseignants de mathématiques, il est donc bon d'en examiner les bases.

Pour commencer, regardons dans un dictionnaire de langue les significations des termes *simuler* et *simulation*. Dans le langage courant, pour le *Petit Robert* (2000), *simuler* veut dire :

« 1. Faire paraître comme réel, effectif (ce qui ne l'est pas) »,

et la *simulation* est :

« 1. Fait de simuler (un acte juridique), de déguiser un acte sous l'apparence d'un autre. 2. Action de simuler ».

Le mot *simulation* ne se trouve pas dans le *Petit Larousse Illustré* de 1972, ce mot n'était pas à cette époque d'usage courant. Dans le langage scientifique et technique, les mots *simulation* et *simuler* ont pris un sens plus spécifique : la *simulation* (mil. XX^{e}) est la :

« représentation du comportement de systèmes physiques (par des calculateurs analogiques, numériques, etc.) en simulant par des signaux appropriés les grandeurs réelles. »

et *simuler* c'est :

« reproduire à l'aide d'un système informatique les caractéristiques et l'évolution (d'un processus) » ;

les corrélats de ces termes étant *modéliser* et *modélisation* :

« Mise en équation d'un phénomène complexe permettant d'en prévoir les évolutions ».

Ces citations du dictionnaire *Le Petit Robert* (2000) s'appuient essentiellement sur les techniques modernes (l'informatique) de la fin du XX^{e} siècle pour caractériser les mots *simulation* et *simuler*.

Notons encore un emploi devenu courant du mot *simulation* dans le domaine économique : en allant dans une banque pour demander un prêt, par exemple pour acheter un appartement ou une maison, les acquéreurs potentiels se voient proposer par le conseiller financier une *simulation du prêt*, c'est-à-dire une application à leur cas d'un modèle d'amortissement d'emprunt : durée du prêt, contraintes de mensualités maximum, etc.

Citons aussi la définition que donne *Encyclopædia Universalis*, Vol.14, 1968 (7^{e} publication, 1976) :

« *La simulation est l'expérimentation sur un modèle. C'est une procédure de recherche scientifique qui consiste à réaliser une reproduction artificielle (*modèle*) du phénomène que l'on désire étudier, à observer le comportement de cette reproduction lorsqu'on fait varier expérimentalement les actions que l'on peut*

exercer sur celle-ci, et à en induire ce qui se passerait dans la réalité sous l'influence d'actions analogues. »

La *modélisation* est une description simplifiée et abstraite de certains aspects de la réalité en termes d'éléments variables et de relations entre ces éléments, et la *simulation* consiste à faire varier ces éléments pour observer ce qui se passe. Le modèle n'est pas nécessairement un modèle mathématique, il peut être un modèle réduit, ou maquette, comme par exemple en architecture pour étudier l'esthétique d'un bâtiment, ses fonctionnalités, son intégration dans le tissu urbain existant ; pour la construction de navires ou de digues, les bassins d'essais pour l'étude de la carène ou les conséquences de l'implantation d'une digue sur l'environnement, etc.

La première étape d'une simulation est donc la modélisation du phénomène étudié, la deuxième étape (dans le cas de systèmes déterministes) étant l'obtention de réalisations à partir de ce modèle, en faisant varier certaines grandeurs en leur donnant des valeurs numériques judicieusement choisies ou en faisant intervenir le *hasard* (méthodes de Monte-Carlo).

Ces définitions du terme *simulation* concernent essentiellement des phénomènes déterministes. Une démarche analogue est utilisée pour étudier le comportement de systèmes stochastiques formés de plusieurs éléments, pour lesquels on peut modéliser la distribution des réponses de chaque élément et les interactions entre éléments, mais où le modèle global du système entier est trop compliqué à analyser et calculer. Des techniques statistiques appropriées permettent ensuite de valider ou non le modèle.

En conclusion, ce n'est donc pas avec une simulation que l'on peut comparer deux modèles (qu'ils soient déterministes ou stochastiques), mais la simulation sert à voir, par l'adéquation des résultats obtenus avec les *observations de la réalité* du phénomène étudié, si le modèle est acceptable ou non pour représenter cette réalité.

2 - Le premier sens de *simulation* probabiliste

En calcul des probabilités, un premier sens pour simuler *une expérience aléatoire*, est de la remplacer par une autre expérience aléatoire, qui a le même modèle probabiliste par rapport à ce qui est étudié - c'est la notion de jeux équivalents que C. HUYGENS avait exprimé de façon implicite : « *je pars de l'hypothèse que dans un jeu la chance qu'on a de gagner quelque chose a une valeur telle que si l'on possède cette valeur on peut se procurer la même chance par un jeu équitable, c'est-à-dire par un jeu qui ne vise au détriment de personne* », - mais qui est plus simple ou facile à réaliser. Par exemple, on veut tirer une personne au hasard parmi un groupe de quatre personnes, le terme *au hasard* signifiant ici que chaque personne doit avoir la même chance d'être tirée. Il existe différents procédés pour effectuer un tel tirage au sort : tirage à la courte paille, désignation à l'aveuglette, à l'aide de comptines enfantines (Pouf ! Ce sera toi qui y seras...), etc., mais on ne peut être certain de réaliser ainsi rigoureusement

l'équiprobabilité. Il est évidemment impensable de mettre chacune des quatre personnes dans des boules identiques (i.e. indiscernables) placées dans une grande urne qu'on secoue pour mélanger ces boules, puis en extraire une. Un moyen plus simple et pratique est de prendre quatre feuilles de papier identiques, de marquer sur chacune d'elles le nom d'une des personnes, de placer ces feuilles pliées de la même façon dans un chapeau ou un sac que l'on secoue pour les mélanger, puis de tirer un des billets à l'aveuglette (c'est le schéma de Jakob BERNOULLI de tirages dans une urne). Le doute sur la réalisation de l'équiprobabilité peut être minimisé par une élaboration soigneuse de l'appareil à fabriquer le hasard, nous mettant en situation de considérer que nous avons bien mis en place une expérience équivalente à celle qui nous était initialement proposée, à savoir le « choix au hasard » d'une personne parmi 4.

On a ainsi une bijection entre les deux expériences *tirer une personne au hasard parmi les quatre* et *tirer un billet au hasard parmi les quatre*, avec le même modèle probabiliste, ce qui permet de faire la réalisation effective d'un *tirage au hasard* dans le groupe de personnes. On peut aussi utiliser les objets usuels à fabriquer du hasard que sont les pièces et les dés. Par exemple avec une pièce (supposée bien équilibrée, sinon voir dans le paragraphe suivant le procédé de FELLER) on associe chacune des quatre personnes à chacun des résultats de deux lancers successifs.

Simulation
d'un jeu de pile ou face !!

Dessin de Christian AFFAGARD

Pince : ça est Pile
Gomme : ça est Face

Cette notion d'équivalence d'expériences aléatoires (par rapport aux propriétés probabilistes) n'est pas, de prime abord, facilement assimilée par les élèves. Est-ce la même chose (c'est-à-dire, obtient-on les mêmes probabilités pour les mêmes observables) si on lance deux dés identiques simultanément, ou l'un après l'autre, ou si on peint un dé en vert et l'autre en rouge ? Ce premier sens (celui de modèle équivalent) est utilisé dans W. FELLER[22] qui au mot *simulation* dans l'index renvoie à un exercice intitulé *Simulation d'une pièce parfaite* avec une pièce biaisée : en lançant deux fois une pièce (biaisée ou non), les événements « (Pile , Face) » et « (Face , Pile) » sont équiprobables, on simule le lancer d'une pièce parfaite en répétant les lancers jusqu'à ce que l'un ou l'autre des deux événements soit obtenu.

Cette équivalence d'expériences aléatoires est à la base des applications du calcul des probabilités et de la simulation probabiliste. En statistique, hormis le cas de l'analyse exhaustive d'un recensement, toutes les applications sont basées sur une simulation en ce premier sens : au départ, il y a tirage *au hasard* d'un échantillon de la population étudiée formée d'individus ou d'objets. Ce tirage d'un

[22] FELLER, *op. cit.*, 3ème éd., p.238.

échantillon aléatoire se fait soit à l'aide d'un dispositif physique ayant le modèle probabiliste désiré ou en utilisant une table de nombres au hasard (dont l'utilisation remonte au début du XX^{e} siècle), qui permet de retrouver, s'il y a besoin, le même échantillon. Il faut bien noter qu'il y a équiprobabilité sur l'ensemble des individus de la population, et non pas sur les valeurs prises par une ou plusieurs caractéristiques étudiées sur ces individus ; un exemple de cette confusion est donné ci-après.

3 - La simulation probabiliste et le premier test d'hypothèse

Une des premières procédures de ce genre est associée à une question démographique posée par John ARBUTHNOT[23] au début du XVIII^{e} siècle : « *Le rapport des sexes à la naissance est-il dû au hasard ou est-il dû à la volonté divine ?* ». L'argument de ARBUTHNOT était le suivant : Si le sexe de l'enfant à naître est déterminé par le *hasard*, c'est-à-dire dans son esprit avec équiprobabilité[24] des naissances mâles et femelles, d'où équiprobabilité pour une année donnée de l'événement « les naissances mâles surpassent les naissances femelles » et de l'événement contraire, alors la probabilité d'observer pendant 82 années successives un excès de naissances mâles sur les naissances femelles est $1/2^{82}$; cet événement étant pratiquement impossible, ARBUTHNOT en conclut qu'il n'y a pas effet du hasard mais intervention de la Divine Providence dans la détermination du sexe d'un enfant à naître.

C'est la première mention de ce qui est appelé maintenant un *test d'hypothèse* ; c'est en quelque sorte une variante probabiliste du raisonnement par l'absurde couramment utilisé en mathématiques. ARBUTHNOT prend comme hypothèse de départ, dite hypothèse nulle H_0, « les deux possibilités, garçon ou fille, pour l'enfant à naître, ont la même chance de se produire », et comme hypothèse alternative « il y a une plus grande chance que l'enfant à naître soit un garçon que ce soit une fille », qu'il interprète comme étant l'expression de la Volonté Divine. A partir de l'hypothèse H_0, par un raisonnement déductif utilisant la théorie des probabilités (en supposant implicitement que les différentes naissances ont des résultats, garçon ou fille, indépendants et que les 82 années comportent à peu près le même nombre de naissances) ARBUTHNOT arrive à la conclusion que le résultat observé, excès de naissances de garçons sur celles des filles pendant 82 années, est

---

[23] ARBUTHNOT, John (1710). An argument for Divine Providence, taken from the constant regularity observ'd in the Birthd of both sexes, *Philosophical Transactions*, **27**, 186-90. Reprod. dans KENDALL, M. et PLACKETT, R. L. (éds) (1977). *Studies in the History of Statistics and Probability*, Vol. II. Ch. Griffin & Co, London.

[24] Cette identification des notions *le phénomène dépend du hasard* et *il y a équiprobabilité* pour la caractéristique étudiée a continué jusqu'à notre époque et est une conception erronée fréquente chez les élèves.

presque impossible, donc presque faux ; il en conclut que l'hypothèse de départ, H_0, est presque fausse.

Nicolas BERNOULLI[25], et après lui P.-S. LAPLACE[26], répondaient par une application de méthodes du calcul des probabilités. L'argument de Nicolas BERNOULLI est d'imaginer, pour simuler l'expérience des naissances à Londres dans une année donnée, le jet de 14 000 dés [nombre milieu des naissances annuelles à Londres sur la période 1629-1710] à 35 faces chacun, dont 18 soient blanches (garçon) & 17 noires (fille)[27], puis de montrer (par une analyse analogue à celle de son oncle Jacques) qu'il y a une grande probabilité (« *il y a beaucoup à parier* ») que l'écart entre le nombre des mâles et 7 200 est plus petit que 163 (écart dans l'année où « *le nombre des filles a été le plus proche de celui des garçons* »), et « *qu'en 82 fois le nombre des mâles ne tombera pas trois fois hors de ces limites* », et il compare alors au « *Catalogue des Enfans mâles et femelles nés à Londres depuis 1629 jusqu'à 1710* » pour valider son modèle.

Ceci est un exemple d'une application importante de la statistique, dans de nombreux domaines, qui concerne la prise de décision sur la distribution d'un caractère sur les individus d'une population à partir de l'observation d'un échantillon aléatoire de cette population. Ce genre de question relève de la statistique inférentielle, et plus particulièrement de la théorie des tests. On peut citer dans le domaine industriel le contrôle de qualité en fabrication, le contrôle de réception de marchandises, etc. La statistique mathématique, sous certaines hypothèses de répartition dans la population, et pour un échantillon aléatoire, détermine les risques associés à une décision et les régions de rejet et d'acceptation pour un test d'hypothèse.

4 - Le deuxième sens de *simulation* en calcul des probabilités

Le deuxième sens des termes *simuler* et *simulation*, celui du langage scientifique et technique (expérimentation sur un phénomène aléatoire), est celui qui nous intéresse car celui-ci est la nouveauté introduite dans les programmes de mathématiques du lycée : utilisation de la simulation comme sensibilisation et illustration pédagogique de phénomènes stochastiques dont le modèle est connu (au moins du professeur). La démarche est la suivante : déterminer le modèle probabiliste du phénomène, construire alors une expérience aléatoire facilement

---

[25] Lettres de N. BERNOULLI à M. de M. dans MONTMORT (*op. cit.*) (1713), p. 373-4 et 388.

[26] LAPLACE, dans son mémoire de 1786, *Sur les naissances, les mariages et les morts à Paris, depuis 1771 jusqu'en 1784,...* trouve le rapport des naissances des garçons à celles des filles comme 25 à 24. Pour l'expérimentation faite à sa demande en 1802 sur 30 départements, le rapport est 22 à 21. Voir son *Essai philosophique* (*op. cit.*), pp.82-87.

[27] N. BERNOULLI assimile ici la probabilité pour chacune des naissances à la fréquence des naissances de garçon et de fille observée sur l'ensemble des 82 années. Le rapport de 18 faces blanches sur 35 est le rapport entier le plus proche de la fréquence des naissances de garçons.

réalisable de même modèle probabiliste que le phénomène étudié, puis effectuer des réalisations de cette dernière expérience aléatoire. On peut noter que le terme *simulation* est aussi utilisé en un sens légèrement plus restreint comme *une* réalisation de l'expérience, on dira alors *faire des simulations de...*

Cependant, faire des simulations pour conduire une expérimentation numérique reproduisant fictivement une expérience aléatoire réelle, n'aboutit pas nécessairement à la construction rapide d'un bon modèle et, pour ce qui nous concerne dans l'enseignement, la mise en place d'outils conceptuels par les élèves. On peut citer un exemple qui s'est produit dans l'école biométrique anglaise à la fin du 19e siècle. WELDON, un biologiste, et PEARSON avaient été impressionnés par le livre *Natural Inheritance* de 1889 de GALTON, et WELDON a poursuivi l'essai de validation statistique de la théorie de l'évolution de DARWIN entamée par GALTON. Il a entrepris (en 1893-4) de vastes lancers de dés pour étudier la variabilité dans les échantillons, afin de la comparer à celle qu'il obtenait pour des observations sur des crabes, et de déterminer si les variations observées sur les crabes pouvaient être dues au hasard ou étaient l'indice d'une évolution dans l'espèce. Un employé de bureau avait fait quelques milliers de lancers de 12 dés et ses résultats ne correspondaient pas bien au modèle théorique d'une loi binomiale. Il y a eu un débat entre GALTON, WELDON, PEARSON et d'autres pour dire si les résultats étaient ou non en adéquation avec le modèle théorique ou si il y avait eu des erreurs de transcription des résultats. C'est le problème bien connu maintenant de l'ajustement d'une distribution observée à une distribution théorique, mais le test du χ^2 n'a été inventé par PEARSON que quelques années plus tard, et en relation à la théorie de la corrélation multiple et non concernant ce problème sur des simulations par lancers de dés.

On peut remarquer à propos de cette expérimentation de WELDON, qu'en dehors des erreurs de transcription des résultats, il peut se faire que les objets utilisés pour les expériences ne vérifient pas exactement les spécifications du modèle supposé, très souvent l'uniformité, c'est-à-dire l'équiprobabilité des différents résultats possibles en nombre fini. LAPLACE[28] l'avait déjà noté :

« souvent entre les faces d'un dé, qui semble parfaitement cube, il existe une inégalité de pente très sensible, en sorte que, sur un fort grand nombre de coups, une des faces arrive plus souvent que l'autre... »

et a regardé l'influence sur l'espérance de gain dans un jeu. On peut citer aussi l'anecdote de KENDALL : lors de tirages répétés avec des jetons de diverses couleurs, on s'était aperçu qu'une couleur ne sortait pas aussi fréquemment qu'elle aurait dû selon le modèle d'équiprobabilité. Un examen attentif des jetons de la

[28] LAPLACE, P. S. (1774), Mémoire sur la probabilité des causes par les événements, *Oeuvres Complètes*, tome 8, Gauthier-Villars, 1889.

couleur en cause comparés aux autres a montré que la peinture de cette couleur-là rendait le jeton légèrement glissant, et par suite moins facile à saisir.

Cette démarche de WELDON est un des premiers exemples de la simulation statistique telle qu'elle est conçue aujourd'hui ; lorsque les combinaisons sont trop complexes pour calculer facilement les probabilités associées aux événements élémentaires du modèle probabiliste, on peut obtenir des approximations de ces valeurs par utilisation du théorème de Jakob BERNOULLI : avec des réalisations ou simulations nombreuses, la fréquence de chaque cas élémentaire est proche de sa probabilité (ou plus précisément la probabilité que l'écart entre la fréquence observée et la vraie valeur dépasse une quantité donnée devient très petite quand le nombre de réalisations devient grand). Ces réalisations peuvent être obtenues à partir d'objets physiques (pièces, dés, boules dans une urne, roue du type loterie, etc.) ou des tables de nombres au hasard (début du XX^{e} siècle) ou de nos jours avec des générateurs de nombres pseudo-aléatoires. Tous ces procédés ont pour modèle une distribution uniforme sur un certain ensemble discret associé à l'expérience.

5 - La méthode de Monte-Carlo

Ce type de démarche est connue aujourd'hui sous le nom de *méthode de Monte-Carlo*[29] ; c'est l'étude, à l'aide de simulations, de certaines propriétés ou du comportement d'un système trop complexe pour être analysé complètement ou dont la résolution du modèle mathématique demande beaucoup trop de calculs, que ce système soit associé à un phénomène aléatoire ou déterministe. La connaissance apportée par cette expérimentation artificielle peut permettre d'affiner le modèle mathématique ou indiquer une analogie avec un problème déjà étudié dans un autre domaine.

L'idée de base repose sur une application de la loi forte des grands nombres : si (X_n) est une suite de variables aléatoires indépendantes de même loi qu'une variable X telle que $E(|X|) < \infty$, alors la moyenne $\frac{1}{n}\sum_{k=1}^{n} X_k$ tend p.s. (presque sûrement) vers E(X) quand *n* tend vers l'infini.

[29] Le nom de Monte-Carlo vient du fait que à la fin du XIX^{e} siècle, des journaux (*La revue de Monte-Carlo*) publiaient les numéros sortis aux tables de roulette du casino de Monte-Carlo, qui passaient pour être les meilleures en Europe. Ces tirages ont été étudiés d'un point de vue statistique, en particulier par Karl PEARSON (Science and Monte-Carlo, *Fortnightly Review*, 1894) qui a plus tard essayé dessus son test d'ajustement du χ^2 et a trouvé que les tirages des roulettes de Monte-Carlo se conformaient de façon tout à fait satisfaisante à une distribution uniforme. Ils étaient utilisés comme nombres au hasard avant la publication de tables spécifiques.

On peut trouver dans le livre, petit en taille mais dense, de HAMMERSLEY et HANDSCOMB[30], des méthodes de Monte-Carlo et des applications, en particulier pour des valeurs numériques d'intégrales.

La méthode de Riemann d'intégration d'une fonction *f* réelle intégrable sur un intervalle I borné consiste à découper cet intervalle I en un nombre fini de segments de même longueur, pour chaque segment à calculer la valeur de la fonction en un point de ce segment (en général le milieu ou une extrémité), d'approximer l'aire élémentaire pour ce segment par l'aire du rectangle correspondant, produit de la longueur de ce segment par la hauteur égale à la valeur de la fonction, puis à faire la somme de toutes ces aires élémentaires. L'utilisation de la méthode de Monte-Carlo pour une approximation de la valeur de l'intégrale est du même genre, mais au lieu de prendre des points déterminés sur l'intervalle I, on prend des réalisations (par exemple, avec un générateur de nombres pseudo-aléatoires) de variables aléatoires (X_n) indépendantes et équidistribuées sur cet intervalle I, alors selon la loi forte des grands nombres (valable ici d'après le théorème de Kolmogorov), la moyenne arithmétique $\overline{f(X)_n} = \frac{1}{n}\sum_{k=1}^{n} f(X_k)$ tend p.s. vers $E(f(X))$ et $E(f(X)) = \frac{1}{\lambda(I)}\int_I f(x)\,dx$, où $\lambda(I)$ est la longueur de l'intervalle I. Cette procédure se généralise sans peine au cas d'une intégrale multiple sur un domaine simple borné, dont on connaît la mesure. Si en outre la fonction *f* est de carré intégrable, on a une évaluation de l'erreur par la variance d'échantillon $\frac{1}{n-1}\sum_{k=1}^{n}[f(X_k) - \overline{f(X)_n}]^2$ et puisque le Théorème-Limite Central est applicable pour *n* assez grand, on peut calculer un intervalle de confiance pour la valeur de l'intégrale.

Cette façon de calculer une valeur approchée d'une intégrale dans le cas à une dimension n'est qu'une illustration de la méthode de Monte-Carlo, car il existe de nombreuses méthodes de quadrature en analyse numérique traditionnelle qui donnent des résultats plus précis et plus rapidement, donc plus efficacement. Ce n'est plus le cas quand le nombre de dimensions est plus élevé, où une méthode de Monte-Carlo peut être avantageuse.

L'inférence statistique qui, en gros, étudie comment les valeurs de caractéristiques mesurées sur les individus d'un échantillon tiré d'une population

---

[30] HAMMERSLEY, J. M. et HANDSCOMB, D. C., *Les méthodes de Monte-Carlo* (traduit de l'anglais). Dunod, Paris, 1967 ; qui indiquent : « *le nom et le développement systématique des méthodes de Monte-Carlo datent de 1944 environ* » ... Leur « *usage comme outil de recherche est né du travail sur la bombe atomique durant la deuxième guerre mondiale. Ce travail entraînait une simulation directe des problèmes probabilistes de la diffusion aléatoire des neutrons dans le matériau fissile* » (pp. 8, 10).

donnée permettent de connaître de façon plus ou moins approchée la distribution de ces caractéristiques sur la population toute entière, a besoin pour ce faire de connaître le comportement de l'échantillon. Le modèle le plus simple en théorie de l'échantillonnage est celui du tirage au hasard d'un échantillon avec remise, c'est-à-dire la répétition d'épreuves identiques et indépendantes ; c'est le modèle de base de la théorie de la décision statistique : tests, intervalles de confiance,.... D'autres modèles d'échantillonnage seront vus à propos des sondages.

Sous de bonnes hypothèses concernant la distribution des caractéristiques dans la population mère (en général, distribution normale ou exponentielle ou gamma dans le cas continu, binomiale, géométrique, de Poisson dans le cas discret), on peut déterminer les distributions théoriques de certaines statistiques d'échantillonnage, moyenne, variance, fonction de répartition empirique. Lorsque les caractéristiques étudiées dans la population n'ont pas une distribution simple, ou que les méthodes mathématiques sont insuffisantes ou trop complexes à mettre en œuvre pour obtenir ces distributions de statistiques d'échantillonnage, on peut s'aider de la méthode expérimentale, ici des réalisations ou simulations d'expériences aléatoires. Cela a été le cas au début du XX^{e} siècle pour la détermination de la distribution du coefficient de corrélation empirique entre deux variables : STUDENT (W.-S. GOSSET)[31] a effectué des expériences d'échantillonnage (des simulations) pour de petits échantillons d'une population normale ($n = 4$ et $n = 8$ avec un coefficient de corrélation $R = 0$ et $R = 0{,}66$) et a conjecturé, à l'aide des courbes de K. PEARSON, la forme de la densité de probabilité du coefficient de corrélation empirique à partir de la distribution d'échantillonnage empirique (distribution uniforme sur $[-1 , +1]$ pour le cas $n = 4$, $R = 0$)[32].

6 - La fabrication du hasard pour une simulation

Le premier problème auquel on est confronté est donc : comment obtenir un échantillon tiré au hasard avec remise d'une variable de distribution donnée sur la population étudiée ? La solution est d'utiliser une expérience aléatoire facilement manipulable qui a le même modèle théorique que celui d'un tirage dans la population (i.e. faire une simulation). Un résultat du calcul des probabilités dit qu'une variable aléatoire de loi donnée peut être construite à partir d'une variable uniforme, construction présentée dans la section suivante.

Le problème de base de l'échantillonnage est ainsi ramené à la réalisation d'une répétition de variables aléatoires indépendantes de même loi uniforme (discrète ou continue sur [0 , 1]), dont on peut tirer la valeur de réalisation d'un échantillon de la

---

[31] STUDENT (1908), Probable error of a correlation coefficient. *Biometrika*, **6**, 302.

[32] D'après E.-S. PEARSON. William Sealy GOSSET, 1876-1937. "Student" as a statistician. *Biometrika*, **30** (1939), reproduit dans *Studies in the History of Statistics and Probability*, eds. E.-S. Pearson et M.-G. Kendall, Ch. Griffin, Londres, 1970.

variable (de loi quelconque fixée) du modèle étudié. Dans la pratique, étant donné que le mesurage d'une grandeur continue ne peut être fait qu'avec un nombre fini de chiffres (qui indique la précision), le phénomène sera simulé à l'aide de répétitions indépendantes d'une épreuve aléatoire ayant un nombre fixé b d'issues équiprobables numérotées par des entiers de 1 à b (ou de 0 à $b-1$), donnant une suite $(a_1, a_2, ...)$ et en prenant un nombre suffisant des a_i pour obtenir la précision voulue par l'intermédiaire de la loi du phénomène. Une telle suite de résultats a_i est appelée une *suite aléatoire en base b* et les a_i de 0 à $b-1$ sont les *chiffres* ($b=2$, binaire, $b=10$, décimal, etc.).

Nous avons vu précédemment la difficulté d'obtenir un dispositif de production du hasard qui fournisse des résultats conformes à l'idée d'une épreuve aléatoire à issues équiprobables. Par exemple, la probabilité d'avoir « Pile » avec une pièce *réelle* sera peut-être très proche de 1/2, mais ne sera pas *exactement* 1/2. Dans le premier tiers du XX^e^ siècle, Karl PEARSON a consacré une partie de son temps à construire – avec l'aide de ses collaborateurs au Laboratoire de Biométrie à University College, London – et à publier des tables numériques et des abaques de fonctions intervenant en statistique (dans la revue *Biometrika* qu'il a créée avec WELDON en 1901 et e.g., *Tables of the Incomplete Beta-function*, 1934) pour faciliter l'application des outils statistiques dans les domaines industriels et agronomiques. Dans le même ordre d'idée, une table de nombres au hasard a été publiée par TIPPETT[33] en 1927, construite à partir d'un dispositif physique soigneusement fabriqué pour assurer, autant que possible, une distribution uniforme et des répétitions indépendantes. L'usage d'une telle table de nombres au hasard permet de reprendre la même série de nombres pour vérifier les calculs ou comparer deux modèles du phénomène étudié (ou comparer deux modèles à la réalité) par rapport à un même ensemble de valeurs, ce que ne permet pas une expérience physique dont les résultats ont une probabilité très petite de se reproduire à l'identique. D'autres tables de nombres au hasard ont été construites sur des principes physiques[34], la dernière en date étant celle de la RAND Corporation[35], construite par écrêtage d'un bruit de fond.

Depuis l'avènement des calculateurs électroniques dans la seconde moitié du XX^e^ siècle, en raison du coût en temps de saisie et en capacité de stockage pour une table de nombres au hasard, il a été préféré un autre procédé, celui des générateurs de nombres pseudo-aléatoires, dont une étude est faite par B. PARZYSZ dans le chapitre suivant.

---

[33] TIPPETT, L. H. C. (1927), *Random Sampling Numbers*, Cambridge.

[34] Par exemple, M. G. KENDALL and B. BABINGTON SMITH (1939). *Tables of random sampling numbers*. Tracts for computers, **24**, Cambridge University Press.

[35] RAND Corporation (1955). *A million Random Digits with 100,000 Normal Deviates*, The Free Press, Illinois.

Une autre possibilité d'obtention de suites aléatoires a commencé à être exploré au début du XX^e^ siècle avec le mathématicien français Emile BOREL. Au carrefour de la théorie de la mesure, de l'arithmétique et de la théorie des probabilités, BOREL[36] a étudié la classe des nombres dont les *chiffres* successifs dans un développement en base *b* possèdent les propriétés attendues d'une suite aléatoire, c'est-à-dire est semblable au résultat de répétitions indépendantes d'une épreuve aléatoire ayant des issues possibles équiprobables numérotées de 0 à $b-1$. Par application de la loi des grands nombres, chaque *chiffre* doit avoir une fréquence qui tend vers $1/b$ et tout groupement donné de *k chiffres* une fréquence qui tend vers $1/kb$, c'est-à-dire qu'il y a équirépartition des *chiffres* et des groupements de chiffres de longueur fixée. BOREL a appelé un nombre possédant une telle propriété *nombre complètement normal en base b* et *absolument normal* un nombre normal en toute base, et il a montré que sur l'intervalle [0 , 1] la mesure (de Borel) de l'ensemble des nombres absolument normaux est 1. Les décimales d'un nombre complètement normal devraient former une suite qui ressemble à une suite aléatoire et, *a priori*, un nombre *choisi au hasard* sur l'intervalle [0 , 1] est presque sûrement un nombre absolument normal. On pensait donc à cette époque que la suite des *chiffres* d'un nombre complètement normal en base *b* permettrait de simuler une suite de tirages indépendants d'une épreuve à *b* issues équiprobables, le tout était d'en trouver un. Cependant, cette définition d'un nombre normal donnée par BOREL est basée sur le comportement asymptotique des *chiffres* du nombre en base *b*. BOREL indiquait d'ailleurs :

« *La question de savoir si un nombre particulier x, normal par rapport à la base 10, est absolument normal, ou même simplement normal par rapport à une autre base, telle que 2, est une question difficile, et le plus souvent insoluble* ».

Cette question : « *un nombre non rationnel donné de [0, 1] est-il normal, absolument normal ?* » reste encore ouverte ; par exemple, bien que la suite des décimales connues de π a toutes les apparences d'une suite aléatoire[37], on ne sait pas si $\pi-3$ ou $1/\pi$ est un nombre normal. Une autre découverte, posant à nouveau la question : *qu'est-ce qu'une suite aléatoire ?*, a été faite par CHAMPERNOWNE en 1933[38] ; il montre que le nombre dont les décimales sont obtenues en écrivant à la suite les entiers successifs est un nombre normal, et pourtant sa construction est tout à fait prévisible. Il a même été montré en 1940 que ce nombre est absolument normal.

Ce dernier exemple met en évidence ici l'absence d'une caractéristique attendue d'une *suite aléatoire*, qui est l'imprévisibilité des chiffres suivants, ayant déjà

---

[36] BOREL, Émile (1926). *Traité du calcul des probabilités et de ses applications*, Tome 2, fasc. 1, *Applications à l'Arithmétique et à la Théorie des fonctions*. Gauthier-Villars, Paris.

[37] Voir l'article de GUILBAUD dans *Le nombre* π, ADCS, 1992.

[38] CHAMPERNOWNE, D. G. (1933) : The construction of decimals normal in the scale of ten. *J. London Math. Soc*. **8**, 254-60.

observé un nombre fini des chiffres de la suite. Cette question a été reprise à la fin du XX^{e} siècle par les logiciens sous le concept de complexité algorithmique. Bien sûr une suite formée de n chiffres 1 est le début d'une suite aléatoire possible, mais sa particularité va à l'encontre de notre intuition de l'aléatoire. Pour les logiciens, une suite est *aléatoire* si les chiffres successifs sont les résultats de répétitions indépendantes d'une épreuve aléatoire, c'est-à-dire que la connaissance des n premiers termes de la suite n'apporte aucune information sur le terme de rang $(n+1)$, et donc sa description algorithmique nécessite autant de nombres que la suite en contient[39]. Cependant, ces travaux n'ont pas encore débouché sur des applications pratiques et la simulation numérique s'effectue maintenant à l'aide de générateurs de nombres pseudo-aléatoires, développés dans la seconde moitié du XX^{e} siècle par les arithméticiens et étudiés dans le chapitre suivant par B. PARZYSZ. Ces générateurs fournissent une bonne simulation d'une loi uniforme discrète. Regardons maintenant comment passer à la simulation d'une loi quelconque.

7 - Quantiles d'une fonction de répartition et simulation d'une distribution quelconque

Soit donc une variable aléatoire X, mesure numérique d'une caractéristique sur les individus d'une population, dont on s'est fixé le modèle théorique de distribution dans la population par une fonction de répartition F, c'est-à-dire $F(x) = P[X \leq x]$, ou bien F est la fonction cumulative des fréquences pour un caractère X sur une population, $F(x)$ est la fréquence des individus i vérifiant $X(i) \leq x$. F est une fonction à valeurs dans [0 , 1] qui est croissante (au sens large), continue à droite, de limite 0 en $-\infty$ et 1 en $+\infty$. La question qui nous intéresse est : ayant un nombre p entre 0 et 1, comment obtenir un nombre x_p vérifiant $F(x_p) = p$. Un tel nombre x_p est appelé *quantile* d'ordre p (ou fractile si p est une fraction) de la distribution de X de fonction de répartition F.

Examinons d'abord le cas le plus simple mathématiquement, celui où F est continue et strictement croissante sur un intervalle $]a\,,b[$, avec $\lim_{x \to a} F(x) = 0$ et $\lim_{x \to b} F(x) = 1$, où $a < b$, a est fini ou $-\infty$, b est fini ou $+\infty$ (cas d'une loi de probabilité continue). F est donc une bijection de $]a\,,b[$ sur $]0\,,1[$ et admet une fonction réciproque F^{-1} continue qui, à chaque valeur de p dans $]0\,,1[$ fait correspondre un nombre unique x_p vérifiant $F(x_p) = p$. La valeur x_p peut être obtenue graphiquement. Ayant tracé le graphe de la fonction F, une droite horizontale d'ordonnée p coupe le graphe de F en un point unique dont l'abscisse est x_p. Le graphe de cette fonction F^{-1} s'obtient à partir du graphe de F par symétrie

[39] Voir, e.g., CHAITIN, G. : Les suites aléatoires, dans *Pour la Science*, dossier *Le Hasard*, 1996 ; ou *La Recherche*, dossier *L'univers des nombres*, 1999.

par rapport à la diagonale d'équation $x = y$ dans un système d'axes orthonormés, avec échange entre axe des abscisses et axe des ordonnées.

Lorsque la distribution est discrète, ce qui est le cas pour la plupart des lois de probabilité étudiées au lycée et le cas de toutes les distributions statistiques observées, la détermination de quelques quantiles (ou ici fractiles) a été abordée au collège en classe de troisième pour la valeur particulière 1/2 qui correspond à la médiane et au lycée pour les valeurs 1/4 et 3/4 qui correspondent aux quartiles, voire même les déciles dans les sections économiques. La médiane et les quartiles sont utilisés pour la construction des boîtes de dispersion (ou boîtes à pattes), qui sont au programme de la classe de première.

Pour une distribution discrète, étudions la détermination graphique des quantiles. La fonction de répartition est une fonction croissante (au sens large) en escalier, le graphe de F est composé de segments horizontaux avec des sauts aux valeurs possibles (variable aléatoire) ou observées (statistique). Joignons les segments horizontaux par un segment vertical pour chaque valeur où il y a saut. On obtient ainsi une courbe continue (appelée courbe cumulative) allant de l'ordonnée 0 pour les valeurs de x inférieures à la plus petite des valeurs possibles ou observées jusqu'à l'ordonnée 1 pour les valeurs de x supérieures à la plus grande des valeurs possibles ou observées. Toute droite horizontale d'ordonnée p, $0 < p < 1$, coupe la courbe cumulative. Si c'est en un seul point (intersection avec un segment vertical), l'abscisse x_p correspondante est le quantile d'ordre p. Si c'est sur un segment horizontal, les valeurs correspondantes de x forment un intervalle $[a_p , b_p]$: il y a une infinité de points possibles comme quantile d'ordre p. Une première possibilité est d'appeler quantile d'ordre p toute valeur de cet intervalle $[a_p , b_p]$. Une seconde possibilité est de prendre une valeur intermédiaire dans l'intervalle ; par exemple, la norme AFNOR préconise de prendre le milieu de l'intervalle médian comme valeur médiane, les tableurs de type EXCEL® donnent une valeur $a_p + k_p b_p$ où k_p dépend de p. Une troisième possibilité est de prendre la borne inférieure a_p de l'intervalle ; ceci résulte d'un point de vue mathématique[40].

Dans le cas d'une distribution discrète, la fonction de répartition F n'est pas bijective et n'a donc pas de fonction inverse au sens ordinaire. On peut procéder de façon heuristique sur la courbe cumulative par analogie aux graphes des fonctions F et F^{-1} lorsque F est bijective. On prend la symétrique C* de la courbe cumulative C par rapport à la diagonale. Cette courbe C* est encore une courbe en escalier, les segments verticaux (resp. horizontaux) de C correspondent à des segments horizontaux (resp. verticaux) de C*. Les segments horizontaux ouverts de C* forment le graphe d'une fonction définie sur $[0 , 1]$ sauf éventuellement aux abscisses de leurs extrémités. L'axe des abscisses porte les valeurs de p (probabilité ou fréquence) et l'axe des ordonnées porte les valeurs de la variable. Cette fonction,

[40] Cf. l'article de J.-C. GIRARD *Quartiles, déciles et tutti déciles* dans ce même volume.

peut être entièrement définie comme l'*inverse généralisée* au sens de LÉVY[41], aussi notée F^{-1}. Paul LÉVY a montré qu'en associant à p dans $]0\,,1[$ le nombre $u_p = \min\{x / F(x) \geq p\}$, on définit une fonction F^{-1} qui est croissante, continue à gauche (donc mesurable) et vérifie $F(x) \geq p \Leftrightarrow x \geq u_p = F^{-1}(p)$. Le graphe de F^{-1} est associé à la courbe C* mentionnée ci-dessus. Alors si U est une variable aléatoire uniforme sur $]0\,,1[$, $X = F^{-1}(U)$ est aussi une variable aléatoire et on a $P[X \leq x] = P[U \leq F(x)] = F(x)$, c'est-à-dire que X admet F comme fonction de répartition.

Le problème de la simulation d'un échantillon d'une loi quelconque F est donc résolu si on dispose d'un moyen de calcul effectif de l'inverse (généralisée) de F et si on a un procédé fournissant des réalisations *indépendantes* suivant une loi *effectivement* uniforme.

Par exemple, en admettant que la suite des décimales de π forme une suite de chiffres au hasard (c'est-à-dire, équirépartis et indépendants), si on les prend par paquets de 5 on a un procédé de simulation d'une loi uniforme sur les nombres à 5 décimales de [0 , 1], d'où la simulation d'une loi quelconque dont on connaît la fonction de répartition.

Application à la loi normale centrée réduite :

On considère, par exemple, les 5 premières décimales de π : 14159. Elles peuvent être considérées comme la réalisation d'une variable uniforme sur [0 , 1] : 0,14159. D'après ce qui précède, on en déduit la réalisation d'une variable normale centrée réduite : – 1,0732 (en utilisant soit la table de la fonction de répartition de la loi normale, soit la fonction LOI.NORMALE.STANDARD.INVERSE d'un tableur), et ainsi de suite avec chaque paquet de 5 décimales.

[41] LEVY, Paul (1923). *Calcul des probabilités*, Gauthier-Villars.

Quelques questions à propos des tables et des générateurs aléatoires

Bernard PARZYSZ

Puisque ces mystères nous dépassent, feignons d'en être l'organisateur.
(Jean Cocteau, *Les Mariés de la tour Eiffel*).

Les nouveaux programmes des lycées font une large place à la simulation d'expériences aléatoires à l'aide de moyens divers, tels que table de chiffres au hasard, touche *random* de la calculatrice, fonction *alea* de tableur-grapheur[1]... Pour l'enseignant comme pour l'élève curieux, peut-être n'est-il pas inutile de se poser et, au moins dans une certaine mesure, de répondre à certaines questions qui viennent immédiatement à l'esprit lorsqu'on évoque ce sujet quelque peu nouveau, qui, comme on le verra, débouche sur d'autres domaines du champ mathématique, en particulier l'analyse et l'algèbre.

I - Pourquoi et comment simuler la répétition d'une même expérience aléatoire de façon indépendante ?

La première idée qui vient immédiatement à l'esprit est... de ne pas la simuler, mais de la *réaliser* : lancers de pièces ou de dés, tirages de boules dans une urne, etc. C'est là effectivement la seule méthode actuellement connue pour produire *réellement* du hasard.

Malheureusement, outre qu'un tel procédé devient vite fastidieux, il faut pouvoir s'assurer que les pièces ou les dés ne sont pas truqués (ou, le cas échéant, qu'ils le sont exactement de la façon souhaitée) ou que toutes les boules présentes dans l'urne ont toutes autant de chances d'être tirées et que les tirages sont indépendants ; c'est-à-dire qu'il faut s'assurer de l'adéquation de l'expérience réalisée avec un modèle équiréparti (ou avec tel autre modèle déterminé à l'avance). Mais cette adéquation ne peut être testée - sans même parler de la prouver - qu'*a posteriori*, ce qui oblige à tout recommencer en cas de désaccord. Les tables de chiffres au hasard, comme les générateurs de nombres aléatoires des calculatrices et des ordinateurs, se proposent justement de pallier ces deux

[1] On trouvera dans le deuxième volume de cet ouvrage des articles traitant de l'utilisation de ces générateurs aléatoires pour exploiter en classe des exemples de simulations.

inconvénients. Tout d'abord, bien sûr, en *accélérant* le processus (pas besoin de réalisation matérielle, il suffit au pire de lire les résultats), mais aussi, et surtout, en introduisant *a priori* le modèle souhaité.

Leur finalité commune est de fournir à l'utilisateur une suite de chiffres (ou de nombres) qu'on peut considérer comme une suite de n tirages dont les issues sont censées être équiréparties et indépendantes. Ainsi, les chiffres d'une table sont censés être équirépartis dans l'ensemble $\{0\,,\ldots,9\}$, les nombres fournis par la touche *Random* d'une calculatrice sont censés être équirépartis sur l'intervalle $[0\,;1[$[2]. Autrement dit : chacun des résultats obtenus est supposé être une réalisation d'une variable équidistribuée sur l'ensemble correspondant.

II - Comment le hasard intervient-il dans la fabrication d'une table de chiffres au hasard ou un générateur aléatoire ?

La réponse à cette question est d'une grande simplicité : le hasard n'intervient pas. Tous les procédés actuellement connus sont de type *déterministe*. Il s'agit donc bien de *simuler* le hasard et non de le *produire* ; c'est pourquoi un qualificatif plus adapté serait en fait celui de *pseudo-aléatoire* (qui est d'ailleurs parfois utilisé). Voici par exemple ce que disent les auteurs d'une table de chiffres au hasard :

« *Une suite de nombres pseudo-aléatoires compris entre 0 et 1 a été engendrée par la méthode multiplicative congruente. Etant donné que nous utilisions un ordinateur IBM 7040 calculant sur 36 bits, nous avons utilisé l'équation de récurrence suivante :* $x_{i+1} = x_i \times 5^{13} \mod(2^{35})$*, dans laquelle* x_i *est le i-ème nombre pseudo-aléatoire, le calcul étant fait en entiers.* $\mathrm{Mod}(2^{35})$ *signifie : diviser* $x_i \times 5^{13}$ *par* 2^{35} *et utiliser le reste - et non le quotient - dans la suite des calculs. Pour construire la table, ces nombres ont ensuite été convertis en virgule flottante et divisés par* 2^{35}*, afin d'obtenir des nombres compris entre 0 et 1. Ensuite ceux-ci ont été multipliés par 10 et tronqués.* »

[ROLF & SOKAL 1969, p. 162. Trad. B. P.]

Il s'agit donc de construire une suite récurrente d'entiers *calibrés* (c'est-à-dire ayant tous le même nombre de chiffres), dont la suite des chiffres est censée constituer autant de réalisations indépendantes d'une même expérience aléatoire. L'affirmation déniant au hasard tout rôle dans le domaine des tables nécessite

[2] Précisons un peu cette expression. La machine - calculatrice ou ordinateur - affiche les nombres compris entre 0 et 1 sous une forme décimale calibrée du type « $a_1a_2\ldots a_p$ », où p est fixé ; on peut ainsi obtenir 10^p nombres différents. Lorsqu'on effectue n « tirages », c'est-à-dire lorsqu'on fait afficher une suite de n tels nombres, la distribution de fréquences de ces nombres définit une statistique X_n à 10^p valeurs. On dira que le tirage est « au hasard » si X_n converge vers la loi équirépartie sur ces 10^p valeurs lorsque n tend vers l'infini.

malgré tout d'être - très légèrement - nuancée. Voici en effet ce que l'on trouve dans le texte d'accompagnement de la table précitée :

« L'accès aux chiffres au hasard doit se faire aléatoirement. Choisissez la page par une procédure aléatoire, déterminez les numéros de la rangée et de la colonne de départ en piquant à l'aveuglette avec un crayon et continuez à partir de là d'une façon prédéterminée, soit horizontalement, soit verticalement. »

[ROLF & SOKAL, ibid.]

Voici maintenant la description d'une autre méthode, attribuée à John VON NEUMANN (1903-1957), l'un des pères de la théorie des jeux :

« *Soit un entier p fixé. On construit une suite d'entiers* $(x_n)_{n\geq 1}$ *s'écrivant tous avec 2p chiffres (en base dix), à l'aide de la relation de récurrence suivante : l'écriture de* x_{n+1} *est constituée des 2p chiffres* centraux *de l'écriture de* $(x_n)^2$. »

N.B. : Puisque l'écriture de x_n comporte $2p$ chiffres, celle de son carré en comporte $4p$.

Voici par exemple ce qu'on obtient par ce procédé pour $p = 2$, à partir de $x_1 = 2\,617$:

2617	8486	0121	0146	0213	0453	2052	2107	4394	3072	4371	1056
1151	3248	5495	1950	8025	4006	0480	2304	3084	5110	1121	2566
5843	1406	9768	4138	1230	5129	3066	4003	0240	0576	3317	0024

Une telle méthode présente cependant certains inconvénients. Ainsi, on risque de tomber sur une suite stationnaire nulle ; c'est précisément le cas ici, car les deux nombres suivants de la suite sont 0005 et 0000. Ceci ne risque pas d'arriver avec la méthode utilisée par ROHLF et SOKAL :

En effet, si on avait $0 = x_i = 5^{13}\, x_{i-1} \mod(2^{35})$, alors (puisque 5^{13} et 2^{35} sont premiers entre eux) on aurait $x_{i-1} = 0$; d'où, de proche en proche, $x_1 = 0$. Or, comme on a bien sûr pris $x_1 \neq 0$, ceci est impossible.

Mais, plus généralement, il n'y a que 10^{2p} nombres à $2p$ chiffres différents. Par conséquent, au bout de $10^{2p} - 1$ itérations au plus (et peut-être moins), on retombera nécessairement sur un nombre déjà obtenu et, du fait de la nature déterministe du procédé, la suite sera périodique. Les constructeurs de tables en sont bien sûr parfaitement conscients, et leur parade consiste à choisir le terme x_1 (le *germe*[3] de la suite) de façon à obtenir une période aussi longue que possible (supérieure, en tout cas, à la longueur de la table). Si par exemple, dans le cas présent, la période est effectivement égale à 10^{2p}, la *longueur maximum autorisée* est une suite de $2p \times 10^{2p}$ chiffres aléatoires.

[3] « seed » (graine) en anglais.

III - Comment engendrer une suite de nombres pseudo-aléatoire ?

Certains mathématiciens se sont intéressés aux nombres réels *modulo* 1[4], en cherchant à engendrer une suite $(u_n)_{n\geq 0}$ de tels nombres qui soit équirépartie sur [0 ; 1[, c'est-à-dire telle que, pour tout entier N, la fréquence des éléments de la suite situés dans chacun des N intervalles $\left[\frac{k}{N}; \frac{k+1}{N}\right]$ est à peu près égale à $\frac{1}{N}$. En se plaçant dans le plan complexe, Hermann WEYL a énoncé le critère suivant (1916) :

« *La suite $(u_n)_{n\geq 0}$ est équirépartie modulo 1 si et seulement si, pour tout entier $p > 0$, on a :* $\lim_{N\to\infty} \sum_{k=1}^{N} exp(2i\pi p\, u_k) = 0$. »

Grâce à ce critère, on montre assez facilement que la suite des multiples d'un irrationnel est équirépartie modulo 1, mais que la suite des multiples d'un rationnel ne l'est pas (voir Annexe 1). Il en résulte que, pour obtenir une suite de nombres équirépartie sur [0 ; 1[, il *suffit* de partir d'un irrationnel α et de considérer la suite $(n\alpha)$. Le seul problème est que les ordinateurs ne savent produire que des nombres décimaux, donc rationnels !

La piste des multiples se révélant ainsi sans espoir, on s'est tourné vers la suite des *puissances* d'un nombre modulo 1, c'est-à-dire à la suite (α^n), avec α donné. On a ainsi pu démontrer que, en partant du nombre d'or $\varphi = \frac{1+\sqrt{5}}{2}$, on n'obtient pas une suite équirépartie. On sait également que l'ensemble des réels supérieurs ou égaux à 1 dont la suite des puissances n'est pas équirépartie est de mesure nulle. Mais on est encore incapable d'identifier les éléments de cet ensemble. En particulier, on ne sait toujours pas à quoi s'en tenir pour un rationnel aussi *simple* que $\frac{3}{2}$...

IV - Existe-t-il des procédés courants pour fabriquer une suite de nombres pseudo-aléatoires ?

Même si on est encore incapable d'engendrer à tout coup une suite de nombres pseudo-aléatoires, il n'est pas interdit d'essayer et c'est bien ce que l'on fait dans la pratique.

Une méthode fréquente de génération d'une telle suite consiste à se placer dans le corps $K = \mathbb{Z}/p\mathbb{Z}$ (où p est un *grand* nombre premier) et à y considérer la suite

[4] C'est-à-dire que le nombre réel x est remplacé par son *résidu modulo* 1, c'est-à-dire $x - \text{Ent}(x)$, où Ent(x) est la partie entière de x.

définie par $u_0 \in K$ et la relation de récurrence $u_{n+1} = a\,u_n + b$, où $(a\,;b) \in K^* \times K$ (suite dite arithmético-géométrique).

N. B. : Par convention, un élément de K sera représenté par son unique élément de $\mathbb{Z}$ compris entre 0 et $p-1$.

Puisqu'une telle suite sera nécessairement périodique, le but visé est, on l'a vu, d'obtenir une suite $(u_n)_{n\geq 0}$ qui *boucle* le plus loin possible, c'est-à-dire pour laquelle le plus petit entier N tel que $u_N = u_0$ soit le plus grand possible (l'idéal étant d'avoir $N = p$).

a) - Un moyen simple consisterait à prendre $a = 1$ et b premier avec p. On a alors, pour tout n, $u_n = u_0 + nb$, d'où $u_n = u_0$ si et seulement si $nb = 0$ modulo p, c'est-à-dire si et seulement si n est multiple de p. Mais une telle suite (arithmétique) est trop *prévisible* pour pouvoir nous être utile. Dans ce qui suit nous considérerons donc $a \neq 1$.

b) - Soit donc $a \neq 1$.

Remarquons tout d'abord que la fonction $x \to ax + b$ admet un point fixe (que nous noterons α). En effet : $\alpha = a\,\alpha + b \Leftrightarrow \alpha = \dfrac{b}{1-a}$.

En posant maintenant (pour tout n) $v_n = u_n - \alpha$, il vient $v_{n+1} = a\,v_n$, c'est-à-dire que la suite $(u_n)_{n\geq 0}$ est géométrique, de raison a.

On a par conséquent, pour tout n, $v_n = a^n\,v_0$, et l'entier N cherché sera en fait l'ordre[5] de l'élément a dans le groupe multiplicatif K^*. Comme cet ordre est un diviseur du cardinal de K^* (égal à $p-1$), nous cherchons donc un élément a dont l'ordre est exactement égal à $p-1$ (c'est ce qu'on appelle une *racine primitive* de p). Un théorème dû à LEGENDRE nous dit qu'il existe $\Phi(p-1)$ telles racines (où Φ est la fonction d'Euler[6]), mais malheureusement, on ne connaît pas actuellement d'algorithme permettant de trouver les racines primitives d'un entier donné.

Les racines primitives de $p-1$ sont-elles rares ?

EULER a démontré[7] que, si n admet $p_1,\ldots, p_k$ comme facteurs premiers, alors on a $\Phi(n) = n \prod_{i=1}^{k}\left(1-\frac{1}{p_i}\right)$. Ainsi, dans le cas qui nous occupe, si on prend un entier au

[5] L'ordre d'un élément k de K^* est le plus petit entier $n > 0$ tel que $k^n = 1$.

[6] $\Phi(n)$ est le nombre d'éléments de $\mathbb{N}^*$ inférieurs à n et premiers avec lui.

[7] On pourra par exemple trouver une démonstration de ce résultat dans [Janvier 2001].

hasard entre 1 et $p-1$, la probabilité qu'il s'agisse d'une racine primitive de p est égale à $\prod_{i=1}^{k}\left(1-\frac{1}{p_i}\right)$, où les p_i sont les facteurs premiers de $p-1$.

Exemple avec une *petite* valeur de p :

Prenons pour p le quatrième nombre de Fermat[8] : $p = 2^{2^4} + 1 = 65\,537$. Ce nombre est premier, et on a $p - 1 = 2^{16}$. Le seul facteur premier de ce nombre est donc 2, d'où la probabilité 0,5 d'obtenir une racine primitive par tirage au hasard dans $\{1 ; \ldots ; 65\,536\}$. Autrement dit : en prenant un nombre entier a au hasard entre 1 et 65 536, on a une chance sur deux que la suite définie dans $K = \mathbb{Z}/65\,537\mathbb{Z}$ par u_0 et la relation de récurrence $u_{n+1} = a\,u_n + b$ ait une période maximale, c'est-à-dire égale à 65 536.

Qu'en est-il pour les calculatrices ?

En 1988, Pierre L'ÉCUYER a présenté une méthode utilisant deux telles suites, dans le but d'augmenter notablement la période [L'ÉCUYER, 1988][9]. Voici de quoi il s'agit (on supposera, par commodité, que les deux suites sont géométriques) :

Soient p et q deux nombres premiers, α une racine primitive de p et β une racine primitive de q, et les deux suites géométriques :

$(u_n)_{n\geq 0}$ définie dans $\mathbb{Z}/p\mathbb{Z}$ par son premier terme u_0 et la relation de récurrence $u_{n+1} = \alpha\,u_n$.

$(v_n)_{n\geq 0}$ définie dans $\mathbb{Z}/q\mathbb{Z}$ par son premier terme v_0 et la relation de récurrence $v_{n+1} = \beta\,v_n$..

On considère alors la suite $(w_n)_{n\geq 0}$ définie dans $\mathbb{Z}/(p-1)\mathbb{Z}$ par la relation de récurrence $w_n = u_n - v_n$.

On a vu que les suites $(u_n)_{n\geq 0}$ et $(v_n)_{n\geq 0}$ ont respectivement pour période $p-1$ et $q-1$. L'ÉCUYER a démontré que la suite $(w_n)_{n\geq 0}$ a pour période le PPCM de ces deux nombres et que, si $(u_n)_{n\geq 0}$ est équirépartie sur $\{1 ; \ldots ; p-1\}$, la suite $(w_n)_{n\geq 0}$ est équirépartie sur $\{0 ; \ldots ; p-2\}$. En outre p et q sont impairs, donc $p-1$ et $q-1$ sont pairs ; s'il est possible de choisir p et q de façon que $\frac{(p-1)}{2}$ et $\frac{(q-1)}{2}$ soient premiers entre eux, la période de la suite $(w_n)_{n\geq 0}$ sera alors $N = \frac{(p-1)\,(q-1)}{2}$.

---

[8] Les nombres de Fermat sont les entiers de la forme $2^{(2^n)} + 1$; certains sont premiers, d'autres non comme par exemple le cinquième qui vaut $2^{32} + 1 = 4\,294\,967\,297 = 641 \times 6\,700\,417$, ainsi que les suivants jusqu'au trentième.

[9] Article communiqué par André GUILLEMOT que je remercie ici pour son aide.

C'est cette méthode qui est par exemple à la base du générateur de la TI 89 et de la TI 92[10]. On a ici $p = 2^{31} - 85 = 2\,147\,483\,563$ et $q = 2^{31} - 249 = 2\,147\,483\,399$ (la calculatrice travaillant sur 32 bits, ces nombres sont aux limites de sa capacité). Ces valeurs répondant à la condition précédente (1 073 741 781 et 1 073 741 699 sont premiers entre eux), la période de la suite $(w_n)_{n\geq 0}$ est de l'ordre de 2^{61}, soit environ 2×10^{18} (puisque $2^{10} \approx 10^3$).

V - Les suites obtenues par ce procédé simulent-elles bien des suites aléatoires ?

Pour l'instant donc (et en attendant les futurs ordinateurs quantiques ?), les générateurs dits *aléatoires* des calculatrices et des ordinateurs ne fournissent, à la demande, qu'une suite *finie* de nombres qui, concaténés, constituent une suite de chiffres qui sera *réputée* aléatoire ; il en va de même pour les tables de chiffres au hasard. Le problème se pose alors de savoir s'il est possible de tester le caractère aléatoire d'une telle suite de chiffres, lorsqu'on est dans l'ignorance du phénomène qui l'a engendrée. La réponse est clairement « non », car *a priori* n'importe quelle suite finie de N chiffres peut être obtenue par la répétition d'une expérience aléatoire dont les dix issues possibles sont 0, 1, 2, ..., 8, 9. On peut tester sa conformité (on la considère comme un échantillon de taille N) au modèle constitué par la répétition d'une expérience aléatoire conduisant à 10 issues ayant la même probabilité $\frac{1}{10}$, comme par exemple le tirage au hasard d'une boule dans une urne contenant 10 boules numérotées de 0 à 9. La suite sera réputée *convenable* si elle n'est pas trop *exotique*, en un sens qui reste à préciser.

Une première possibilité est de s'intéresser à la distribution des fréquences des 10 chiffres. Une *bonne* suite sera alors une suite dans laquelle les fréquences de 0, 1, 2, ..., 8, 9 seront toutes *voisines* de 0,1. Pour tester une telle équirépartition des chiffres sur la suite finie dont on dispose, on peut effectuer un test classique, comme celui du Khi-deux[11].

Rappelons que ce test consiste à calculer la *distance du* χ^2 entre la distribution des fréquences des chiffres de 0 à 9 dans la suite donnée et la loi de probabilité

---

[10] D'après un document inédit d'André GUILLEMOT intitulé *Fonction Random et arithmétique*.

[11] Cf. l'article de Louis-Marie BONNEVAL et Michel HENRY : *Tests d'adéquation à une loi équirépartie et autres tests du Khi-deux,* à la fin de ce volume. Rappelons que le test du Khi-deux repose sur l'utilisation de la « distance » du χ^2 (il s'agit en fait d'une pseudo-distance, car elle n'est pas symétrique), portant sur les probabilités sur {0, …, 9}, autrement dit les suites $(p_0, \ldots, p_9)$ de nombres positifs tels que. $\sum_{j=0}^{9} p_j = 1$.

uniforme $\{\frac{1}{10}, \dots, \frac{1}{10}\}$, puis à comparer cette distance d à une *distance critique* d_c, déterminée par le risque accepté de se tromper quand, d étant inférieure à d_c, on conclut à l'équirépartition des chiffres dans la suite.

Par exemple, si l'événement « obtenir un échantillon aléatoire d'une loi uniforme sur {0 , … , 9} dans lequel la distribution des fréquences est plus éloignée que celle observée » a une probabilité supérieure à 0,05, on considérera que la suite donnée ne s'écarte pas significativement de l'équirépartition, au seuil de signification 0,05.

C'est d'ailleurs ce type de souci qui a conduit les auteurs de certaines tables de chiffres au hasard à *corriger* la suite initialement obtenue, de façon à se rapprocher de la distribution des fréquences de la distribution théorique :

« *Une analyse de la table a révélé que certains chiffres (3 et 7) étaient un peu trop fréquents et que les chiffres 2 et 8 étaient un peu trop rares (quoique restant dans les limites d'une erreur d'échantillonnage normale). Nous avons donc réalisé une transformation pour ajuster la suite de nombres de telle sorte qu'ils soient distribués de façon plus uniforme.* » [ROLF & SOKAL, ibid.]

Cependant, même si le test conclut à la conformité (ou plutôt : s'il ne permet pas de rejeter l'hypothèse de conformité), il sera insuffisant. En effet, la suite $(r_n)_{n\geq 1}$ constituée de répétitions de la suite des 10 chiffres 0, 1, 2,..., 8, 9 est manifestement équirépartie (on trouverait un χ^2 pratiquement égal à 0 pour toute sous-suite finie assez longue), mais personne ne l'accepterait comme une suite *aléatoire*. Il en va de même pour la suite $(s_n)_{n\geq 1}$ constituée de k répétitions du chiffre 0, suivies de k répétitions du chiffre 1, ..., suivies de k répétitions du chiffre 9, etc. La raison en est que, intuitivement, le hasard est perçu comme étant imprévisible et que le caractère répétitif (et donc prédictible) de ces deux suites contredit cette intuition. Le problème devient alors celui de la prise en compte de cette *imprévisibilité*.

Dans ce but, on peut s'intéresser à l'écart entre deux occurrences successives d'un même chiffre. Mais pour cela, il nous faut définir la notion d'équirépartition pour une suite infinie de chiffres. Une suite de *N* chiffres sera alors considérée comme le *début d'une suite infinie*. De façon analogue à ce que nous avons vu pour les réels modulo 1, on peut définir l'équirépartition des chiffres de la façon suivante :

On dit que la suite infinie de chiffres $(x_n)_{n\geq 1}$ est équirépartie *si la fréquence de chacun des dix chiffres dans l'ensemble de ses N premiers termes tend vers 0,1 lorsque N tend vers l'infini :*

$$\forall k \in \{0 , \dots , 9\} \quad \lim_{N \to \infty} \frac{Card\{x_n / n \leq N \text{ et } x_n = k\}}{N} = 0{,}1.$$

Si la suite infinie $(x_n)_{n\geq 1}$ est aléatoire, alors la loi de cet écart (différence de deux rangs d'apparition successifs) est indépendante du chiffre considéré. C'est une loi géométrique de paramètre 0,1 : la probabilité d'observer un écart de longueur l donnée ($l \geq 1$) est égale à $0{,}9^{l-1} \times 0{,}1$.

On est alors conduit à comparer l'ensemble des écarts observés à cette distribution théorique. Si pour tout chiffre k, la distribution des fréquences des écarts entre deux occurrences consécutives de k dans une sous suite finie de taille n converge vers cette loi géométrique, on dira que la suite obtenue est *bien enchevêtrée*[12] [GLAYMANN 1976].

N.B.1 : Ce test, effectué sur les 10 000 premières décimales de π, montre que la répartition est très proche de la répartition théorique.

N.B.2 : Par contre, les suites $(r_n)_{n\geq 1}$ et $(s_n)_{n\geq 1}$ définies plus haut ne sont pas bien enchevêtrées.

Il est une autre question qui se pose à l'utilisateur d'un générateur aléatoire : celle de ses modalités d'utilisation. Nous avons vu que ce qui est initialement engendré par la machine n'est pas une suite de chiffres, mais une suite de N *nombres*, compris entre 0 et 1. Dans une table de chiffres aléatoires comme sur l'écran de la calculatrice ou de l'ordinateur, chaque nombre est tronqué à k décimales ($k \geq 1$) pour fournir la k-suite des chiffres de sa partie décimale. En concaténant ces k-suites, on dispose alors d'une suite (finie) de Nk chiffres.

L'utilisation la plus courante consiste alors à subdiviser cette suite en *blocs* de n chiffres, considérés comme écritures de nombres à n chiffres, en faisant l'hypothèse que ces nombres sont équirépartis sur l'ensemble $\{0, \ldots, 10^n - 1\}$. Par exemple, pour simuler le tirage d'une boule dans une urne contenant 38 % de boules noires et 62 % de boules blanches, on pourra subdiviser la suite en *nombres* de deux chiffres : un nombre compris entre 00 et 37 correspondra au tirage d'une boule noire, et un nombre compris entre 38 et 99 correspondra au tirage d'une boule blanche.

Malheureusement, l'équirépartition de la suite initiale de chiffres ne suffit pas à assurer celle des nombres ainsi obtenus. On peut le voir sur l'exemple des suites $(r_n)_{n\geq 1}$ et $(s_n)_{n\geq 1}$ définies plus haut :

---

[12] Définissons un peu plus précisément cette notion. Soit k un chiffre décimal ($k \in \{0, 1, \ldots, 9\}$). On considère les p+1 premières occurrences du chiffre k dans la suite $(x_n)_{n\geq 1}$, ce qui donne p écarts successifs entre ces occurrences. On obtient ainsi une distribution de p fréquences, définissant une statistique X_p à valeurs dans $\mathbf{N}^*$. On dira que la suite $(x_n)_{n\geq 1}$ est bien enchevêtrée si , pour tout chiffre k, la distribution de X_p converge vers la loi géométrique de paramètre $\frac{1}{10}$ lorsque p tend vers l'infini.

- pour la première, la concaténation 10 à 10 ne donne que l'unique nombre 0123456789 (parmi 10^{10} nombres possibles)
- pour la seconde, la concaténation k à k ne donne que 10 nombres (parmi les 10^k possibles).

On est ainsi amené à renforcer la notion d'équirépartition :

La suite de chiffres $(x_n)_{n\geq 1}$ est dite parfaitement équirépartie *si, pour tout $k \geq 1$, la suite de nombres $(y_n)_{n\geq 1}$ (où y_n est le nombre entier qui s'écrit $\overline{x_{(n-1)k+1}\ldots x_{nk}}$ en base dix) est équirépartie*[13].

Emile BOREL a appelé *nombre normal* un nombre réel de l'intervalle]0 ; 1[dont la suite des décimales est parfaitement équirépartie. On démontre que la probabilité qu'un nombre pris au hasard dans cet intervalle[14] soit normal est égale à 1 (mais ni les nombres rationnels, ni les nombres algébriques ne sont normaux). On démontre également que toute suite *translatée* d'une suite parfaitement équirépartie l'est aussi [HENNEQUIN, *ibid.*]. C'est-à-dire que, si le nombre x est normal, il en est de même du nombre 10^{nx} (pour tout entier n) . On démontre également que la suite des décimales d'un nombre normal est bien enchevêtrée (cf. Annexe 2).

En conséquence, si l'on tire un nombre au hasard dans l'intervalle]0 ; 1[[15], il y a *toutes les chances* pour que la suite de ses décimales puisse être considérée comme aléatoire, de même que toute suite de nombres à n chiffres obtenue par concaténation de chiffres successifs (et ceci en partant de n'importe quel chiffre).

N.B. : La probabilité que le nombre tiré ne soit pas normal est nulle, mais cela ne signifie pas que cet inconvénient ne se produira pas : la probabilité d'obtenir un rationnel en tirant un nombre au hasard dans]0 ; 1[est nulle, elle aussi !

Notre problème serait-il alors résolu ? Plus précisément, suffirait-il de considérer la suite des décimales d'un nombre normal (ou la suite des entiers obtenus par concaténation n à n) pour disposer d'une suite pouvant être considérée comme *aléatoire* ?

---

[13] On peut définir l'équirépartition de la suite des nombres $(y_n)_{n\geq 1}$ de la manière suivante :

pour tout entier $a \in \{0,\ldots, 10^k-1\}$, on a $\lim\limits_{N\to\infty} \dfrac{\text{Card}\{y_n / n \leq N \text{ et } y_n = a\}}{N} = 10^{-k}$.

[14] Il convient ici de préciser le sens de cette expression, qui apparaît entre autres dans le document d'accompagnement des actuels programmes de terminale. Elle signifie en fait qu'on se place dans un modèle probabiliste faisant intervenir une loi continue uniforme sur]0 ; 1[. C'est en pratique ce modèle qui est utilisé pour représenter un tirage "au hasard" dans un ensemble de décimaux "calibrés" (c'est-à-dire avec un nombre de chiffres significatifs fixé), comme c'est le cas pour une calculatrice ou un ordinateur.

[15] Voir note précédente.

Malheureusement il n'en est rien, et en particulier CHAMPERNOWNE exhiba au début du XX^{e} siècle, en guise de contre-exemple, le fameux nombre d'écriture décimale 0,12345678910111213141 5… (obtenu en concaténant les écritures décimales des entiers successifs), qui est normal.

D'autre part, les résultats que nous venons d'obtenir sont uniquement valables *à l'infini*, alors que nous ne pouvons bien évidemment travailler que sur des ensembles *finis* de nombres *tronqués*, dont nous ignorons même s'ils sont normaux (ainsi, on ignore si $\frac{1}{\pi}$ et $\frac{1}{e}$ le sont).

En 1965, P. MARTIN-LÖF proposa d'exploiter l'idée générale suivante : une suite de chiffres sera dite aléatoire si elle ne vérifie aucune propriété *exceptionnelle* qu'on peut *réellement* tester , *exceptionnelle* signifiant que la propriété n'est vérifiée que par une infime partie des suites (c'est-à-dire un ensemble de mesure nulle) et *réellement* signifiant que le test est effectué grâce à un programme informatique avec une précision d'autant plus grande qu'on dispose de plus de chiffres de la suite[16]. Il en résulte que la suite des chiffres d'une suite aléatoire (au sens de MARTIN-LÖF) ne peut être définie par un programme. Comme le dit J. P. DELAHAYE, la *« gravissime inefficacité [de la théorie de MARTIN-LÖF] rend difficile et presque impossible, toute utilisation pratique de ses résultats »*[17]. D'où la démarche qui est actuellement de règle, à savoir utiliser une simulation qui sera éventuellement testée *a posteriori*.

Ainsi utilise-t-on dans ce but des tests divers et variés. Contentons-nous d'en citer deux, parmi les plus courants :

1) - Dans la lignée de ce que nous venons de voir, on peut tester la table en groupant ses chiffres n par n et en contrôlant l'équirépartition des entiers à n chiffres ainsi obtenus dans l'ensemble $\{0, \ldots, 10^n - 1\}$. On effectue ce test pour $n = 1, 2, 3...$

2) - En fixant $n = 5$, on peut tester, non pas la répartition de tous les nombres compris entre 0 et 99 999, mais celle d'une partition de ces nombres, définie en l'occurrence à partir du jeu de poker :

- ceux qui comportent 5 chiffres différents ;
- ceux qui comportent une paire et une seule (deux chiffres identiques) ;
- ceux qui comportent deux paires ;
- ceux qui comportent un brelan (3 fois le même chiffre) ;

---

[16] Pour plus de détails, on pourra se reporter à [DELAHAYE 1999a] chapitre 4 et [DELAHAYE 1999b] chapitre 2.

[17] [DELAHAYE 1999a] p. 39.

- ceux qui comportent un full (un brelan et une paire) ;
- ceux qui comportent un carré (4 fois le même chiffre) ;
- ceux qui comportent un poker (5 fois le même chiffre).

En supposant l'équiprobabilité des 10^5 séquences (ordonnées) de 5 chiffres possibles, la théorie des probabilités fournit les résultats suivants :

Type de séquence	Formule	Probabilité
5 chiffres différents	$\frac{5!}{10^5} \times \binom{10}{5}$	0,3024
une paire	$\frac{10 \times 3!}{10^5} \times \binom{5}{2}\binom{9}{3}$	0,5040
deux paires	$\frac{5 \times 3!}{10^5} \times \binom{3}{2}$	0,1080
Brelan	$\frac{3!}{10^5} \times \binom{5}{2}\binom{10}{3}$	0,0720
Full	$\frac{2!}{10^5} \times \binom{5}{2}\binom{10}{2}$	0,0090
Carré	$\frac{5 \times 2!}{10^5} \times \binom{10}{2}$	0,0045
Poker	$\frac{10}{10^5}$	0,0001

On compare ensuite les résultats obtenus avec les résultats théoriques, par exemple à l'aide d'un χ^2.

VI - … Et pourquoi pas π ?[18]

La suite des décimales du nombre π a, de tout temps, suscité l'intérêt des mathématiciens. Tout d'abord pour essayer de savoir si l'on ne pourrait pas y déceler une périodicité, ce qui se révéla vain dès qu'on eut démontré que π est irrationnel. A défaut, on essaya - et on essaie toujours - d'y repérer des particularités qui permettraient de mieux connaître ce nombre mythique, ce qui impose de rechercher toujours plus de décimales (le 20 septembre 1999, on a dépassé 206 milliards[19]). Ce faisant, on s'est aperçu que, dans cette suite (finie), la

[18] La source des données de ce paragraphe est essentiellement [DELAHAYE 1997].

[19] Yasumasa KANADA, de l'université de Tokyo, a obtenu ce résultat en 37 heures. Il est régulièrement recordman dans ce domaine où excellent aussi les frères CHUDNOVSKY, de l'université de Columbia.

fréquence de chaque chiffre est très voisine de 0,1 : à titre d'exemple, voici ce que donnent les 6 premiers milliards[20] :

Chiffre	Effectif observé	Différence
1	600 033 260	+ 33 260
2	599 999 169	– 831
3	600 000 243	+ 243
4	599 957 439	– 43 561
5	600 017 176	+ 17 176
6	600 016 588	+ 16 588
7	600 009 044	+ 9 044
8	599 987 038	– 12 962
9	600 017 038	+ 17 038
0	599 963 005	– 36 995

De même, grâce au test du poker (voir ci-dessus), on peut observer que, sur les 10 millions de premières décimales, les deux millions de séquences de 5 chiffres obtenues ont des fréquences très proches de la répartition théorique :

Type de séquence	Effectif théorique	Effectif observé	Écart	Écart (%)
5 chiffres différents	604 800	604 976	+ 176	+ 0,0291
1 paire	1 008 000	1 007 151	– 849	– 0,0842
2 paires	216 000	216 520	+ 520	+ 0,2407
brelan	144 000	144 375	+ 375	+ 0,2604
full	18 000	17 891	– 109	–0,6056
carré	9 000	8 887	– 113	– 1,2556
poker	200	200	0	0

En outre, les recherches récentes effectuées sur les décimales de π indiquent qu'il pourrait bien être normal. D'où l'idée d'utiliser comme liste de chiffres *au hasard* la suite des premières décimales de π, obtenue à partir de l'un des algorithmes connus (cf. Annexe 3)… sauf toutefois dans les applications cryptographiques, car on pourrait le *démasquer*, grâce précisément à la suite de ses premières décimales.

VII - Qu'en conclure ?

Il ne s'agit pas pour autant de jeter le bébé avec l'eau du bain en prétextant, soit que le générateur de la calculatrice ou de l'ordinateur n'est pas aléatoire, soit que l'on n'est pas certain que la suite (finie) de chiffres ou de nombres obtenus soit sensiblement équirépartie. Il se trouve que les générateurs actuels fonctionnent

[20] [DELAHAYE 1997] p. 172

assez bien, comme chacun peut s'en rendre compte en les utilisant. En tant qu'enseignants, ce qu'il faut surtout, c'est faire en sorte que les élèves s'en persuadent, afin de tempérer *l'effet boîte noire* inhérent à l'utilisation d'une machine, ou plus généralement d'une procédure sur laquelle l'utilisateur n'a pas prise. Pour cela, la répétition effective d'une même expérience aléatoire (à la main) et la comparaison avec diverses simulations, même si elle reste nécessairement qualitative, est un préalable incontournable, ainsi que la compréhension de *ce qu'on met dans la machine* pour simuler diverses expériences aléatoires. Du point de vue conceptuel, ce dernier point est en effet fondamental, puisque c'est bien un *modèle théorique* qu'on introduit dans la calculatrice ou l'ordinateur pour en simuler des réalisations. Il s'agit en fait de persuader les élèves du fait que la machine constitue un *prolongement* de la main mais ne s'y réduit pas, et que, si on peut en conséquence lui accorder une certaine confiance, celle-ci n'exclut pas la vigilance (pertinence du modèle). Cette phase essentielle de l'apprentissage ne saurait être escamotée, sous peine de compromettre, ultérieurement, l'accès au concept de probabilité.

Références bibliographiques

DELAHAYE, J. P., (1997) : *Le fascinant nombre π*, Paris, Belin, coll. Pour la Science.

DELAHAYE, J. P., (1999a) : *Logique, informatique et paradoxes*, Paris, Belin, coll. Pour la Science, 3ème édition.

DELAHAYE, J. P., (1999b) : *Information, complexité et hasard*, 2ème édition revue, Paris, Hermès Science Publications.

DRESS F. et MENDÈS-FRANCE M., (2001) : La suite des puissances de $\frac{3}{2}$, in *La Recherche* n° 346, pp. 34-37.

ENGEL A., (1990) : *Les certitudes du hasard*, Lyon, Aléas.

GLAYMANN, M., (1976) : L'enchevêtrement des chiffres d'une table de chiffres aléatoires, in *Hasardons-nous*, pp. 125-137, Paris, APMEP, brochure n° 17.

HENNEQUIN, P. L., (1976) : Quelques remarques sur l'article précédent, in *Hasardons-nous*, pp. 139-144, Paris, APMEP, brochure n° 17.

JANVIER, M., (2001) : Les nombres pseudo-aléatoires, in *Des statistiques à la pensée statistique*, pp. 165-187, IREM de Montpellier.

L'ÉCUYER, P., (1988) : Efficient and portable combined random number Generators, in *Communications of the ACM* 31-5, pp. 742-749 + 774.

ROHLF, F. J. & SOKAL R. R., (1969) : *Statistical Tables*, San Francisco, W. H. Freeman & Co.

VAGOST, D., (2000) : *Problèmes d'échantillonnage. Utilisation de la simulation* (texte inédit)

Annexe 1

Proposition :

La suite des multiples d'un irrationnel (resp. rationnel) est équirépartie (resp. n'est pas équirépartie) modulo 1.

Soit α un réel strictement positif, et considérons la suite $(n\alpha)$.

Pour p et N entiers naturels non nuls, posons $Sp(N) = \frac{1}{N} \sum_{k=1}^{N} e^{2i\pi\, pk\ \alpha}$.

On peut encore écrire : $Sp(N) = \frac{1}{N} \sum_{k=1}^{N} \left(e^{2i\pi\ p\alpha}\right)^k$, d'où

- Si $p\alpha$ n'est pas entier : $Sp(N) = \frac{1}{N}\ e^{2i\pi\, p\alpha}\, \frac{1 - e^{2i\pi\, pN\alpha}}{1 - e^{2i\pi\, p\alpha}}$.

- Si $p\alpha$ est entier : $Sp(N) = 1$.

Il en résulte :

(i) si α est irrationnel :

Aucun $p\alpha$ n'est entier, donc $\left|\frac{1 - e^{2i\pi\, pN\alpha}}{1 - e^{2i\pi\, p\alpha}}\right| < \frac{2}{\left|1 - e^{2i\pi\, p\alpha}\right|}$ et $\lim_{N \to \infty} S_p(N) = 0$

quel que soit p. D'après le critère de Weyl, la suite $(n\alpha)$ est équirépartie.

(ii) si α est rationnel :

Il existe un entier p tel que $p\alpha$ soit entier, d'où (pour ce p) $\lim_{N \to \infty} S_p(N) = 1$.

D'après le critère de Weyl, la suite $(n\alpha)$ n'est pas équirépartie.

Annexe 2

Proposition :

Soit une suite $(x_n)_{n\geq 1}$ de chiffres aléatoire. Si la suite de chiffres $(x_n)_{n\geq 1}$ est parfaitement équirépartie, alors elle est bien enchevêtrée.

Avant de démontrer cette proposition, nous allons démontrer le lemme suivant :

Lemme :

Dans les conditions de la proposition ci-dessus, soit $k \in \mathbb{N}^*$ fixé et $\overline{a_1 a_2 \ldots a_k}$ l'écriture décimale d'un nombre entier donné à k chiffres. Soit A l'ensemble des entiers $n \geq 1$ pour lesquels on a $(x_n, \ldots, x_{n+k-1}) = (a_1, \ldots, a_k)$. Alors on a :

$$\lim_{N \to \infty} \frac{\text{Card}\{n / n \in A \text{ et } n \leq N\}}{N} = 10^{-k}.$$

1 - Preuve du lemme :

Notons A_i l'ensemble des indices $n \in A$ tels que $n \equiv i \mod(k)$. On a bien sûr $A = \bigcup_{i=0}^{k-1} A_i$ (réunion disjointe), d'où :

$$\text{Card}\{n / n \in A \text{ et } n \leq N\} = \sum_{i=0}^{k-1} \text{Card}\{n / n \in A_i \text{ et } n \leq N\}.$$

Puisque la suite $(x_n)_{n\geq 1}$ est parfaitement équirépartie, d'après la définition donnée dans l'article, pour tout i la suite des nombres $(I_p)_{p\geq 1}$ (où I_p est le nombre à k chiffres qui s'écrit $\overline{x_{i+1+(p-1)k} \ldots x_{i+pk}}$ en base dix) est équirépartie.

Or, en désignant par $\text{Ent}\left(\frac{N}{k}\right)$ la partie entière de $\frac{N}{k}$, avec les x_n tels que $1 \leq n \leq N$, on peut constituer $\text{Ent}\left(\frac{N}{k}\right) - \varepsilon$ nombres de la suite $(I_p)_{p\geq 1}$, où $\varepsilon = 0$ pour $i \leq N - k\,\text{Ent}\left(\frac{N}{k}\right)$ et $\varepsilon = 1$ pour $i > N - k\,\text{Ent}\left(\frac{N}{k}\right)$.

On a donc $\lim_{N\to\infty} \dfrac{\text{Card}\{n / n \in A_i \text{ et } n \le N\}}{\text{Ent}\left(\dfrac{N}{k}\right) - \varepsilon} = 10^{-k}$, et comme, pour N tendant vers l'infini, $\text{Ent}\left(\dfrac{N}{k}\right) - \varepsilon$ est équivalent à $\dfrac{N}{k}$, on en déduit :

$$\lim_{N\to\infty} \frac{\text{Card}\{n / n \in A_i \text{ et } n \le N\}}{N} = \frac{10^{-k}}{k}.$$

D'où finalement :

$$\lim_{N\to\infty} \frac{\text{Card}\{n / n \in A \text{ et } n \le N\}}{N} = \lim_{N\to\infty} \sum_{i=0}^{k-1} \frac{\text{Card}\{n / n \in A_i \text{ et } n \le N\}}{N} = k\,\frac{10^{-k}}{k} = 10^{-k}.$$

2 - Preuve de la proposition :

Soit m un chiffre donné. Dans la suite aléatoire $(x_n)_{n\ge1}$, on considère les séquences de chiffres apparaissant entre deux occurrences consécutives de m, de la forme $(x_r, x_{r+1}, \ldots, x_{r+e})$ avec $x_r = x_{r+e} = m$, les autres éléments étant différents de m. Pour une telle séquence, on désigne par R_m la variable aléatoire égale à la différence des rangs de ses extrémités.

D'après le lemme, la suite $(x_n)_{n\ge1}$ étant parfaitement équirépartie, la fréquence d'apparition de chaque séquence à $e+1$ chiffres dans les N premiers termes de la suite a pour limite $10^{-(e+1)}$ lorsque N tend vers l'infini. Or, pour m fixé, il y a 9^{e-1} nombres à $e-1$ chiffres distincts de m et donc $10 \times 9^{e-1}$ nombres en faisant varier m de 0 à 9. Une séquence de chiffres d'extrémités m étant prise au hasard, d'après la loi des grands nombres, la probabilité de l'événement $\{R_m = e\}$ est la *limite* des fréquences de telles séquences parmi les N premiers éléments de la suite $(x_n)_{n\ge1}$ quand N tend vers l'infini, soit $\dfrac{9^{e-1}}{10^{e+1}}$.

Quand m varie de 0 à 9, la loi de la variable R égale à l'écart entre deux occurrences consécutives d'un même chiffre, est donc donnée par :

$$\text{P}(R = e) = 10.\frac{9^{e-1}}{10^{e+1}} = 0{,}9^{e-1} \times 0{,}1.$$

On en conclut que R suit une loi géométrique de paramètre 0,1 ; la suite $(x_n)_{n\ge1}$ est donc bien enchevêtrée selon la définition du paragraphe V (note 12).

Annexe 3

Voici, d'après [DELAHAYE 1997], des algorithmes permettant d'obtenir assez rapidement les premières décimales de π :

1°) - <u>Formule de Ramanujan</u> (1910)

Grâce à elle GOSPER calcula, en 1985, 17 millions de décimales. À chaque incrémentation de *n*, on obtient 8 décimales exactes supplémentaires :

$$\pi = \frac{9\,801}{\sqrt{8}} \sum_{n=0}^{\infty} \frac{(4n)!\,(1\,103 + 26\,390\,n)}{(396(n!))^4}$$

2°) - <u>Formule des frères Chudnovsky</u> (1994)

Elle leur a permis de calculer 4 milliards de décimales. À chaque incrémentation de n, on obtient 14 décimales supplémentaires.

$$\pi = 12 \sum_{n=0}^{\infty} \frac{(-1)^n\,(6n)!\,(13\,591\,409 + 545\,140\,134\,n)^{-1}}{(3n)!\,(n!)^3\,640\,320^{3(n+0,5)}}$$

Remarque :

Ces développements en série présentent le défaut majeur de nécessiter beaucoup de calculs, aussi depuis quelques années utilise-t-on plutôt des algorithmes basés sur des suites récurrentes telles que les suivantes.

3°) - <u>Suite de Salamin et Brent</u> (1973)

Issue des travaux de GAUSS sur la moyenne arithmético-géométrique, elle double le nombre de décimales à chaque pas et fournit, avec *seulement* 25 itérations, 45 millions de décimales exactes (mais encore faut-il disposer d'une machine capable de calculer avec 45 millions de décimales !).

$$a_0 = 1 \quad b_0 = \frac{1}{\sqrt{2}},\ c_0 = \frac{1}{2}$$

$$a_n = \frac{1}{2}(a_{n-1} + b_{n-1}) \quad b_n = \sqrt{a_{n-1}\,b_{n-1}} \quad c_n = c_{n-1} - 2^n\,(a_n^2 - b_n^2) \quad p_n = 2\,\frac{a_n^2}{c_n}.$$

4°) - Suite des frères Borwein (1987)

Elle quadruple le nombre de décimales exactes à chaque pas et donne 10 millions de décimales après 16 itérations.

$a_0 = \sqrt{2} - 1 \quad b_0 = 6 - 4\sqrt{2}$

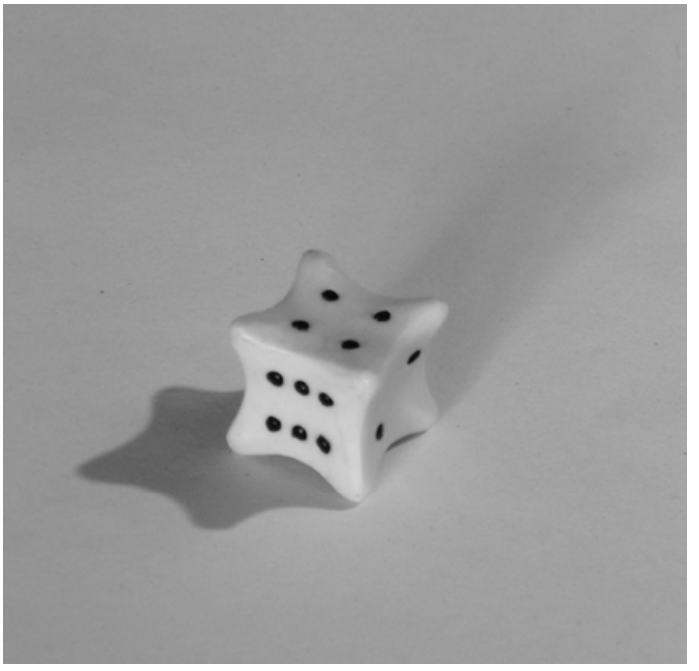
$$a_n = \frac{1 - \left(1 - a_{n-1}{}^4\right)^{1/4}}{1 + \left(1 - a_{n-1}{}^4\right)^{1/4}} \quad b_n = b_{n-1}\,(1 + a_n)^4 - 2^{2k+3}\,a_n\,(a_n{}^2 + a_n + 1) \quad p_n = \frac{1}{b_n}.$$

Pour terminer, mettons à part la formule suivante, surprenante par sa simplicité et trouvée presque *par hasard*, qui possède en base 16 une propriété dont on aimerait bien pouvoir disposer en base 10.

5°) - Formule de Bailey-Borwein-Plouffe (1996)

Très rapide, elle permet de calculer directement le *n*-ième chiffre du développement de π en base 16, sans avoir besoin de connaître les précédents.

$$\pi = \sum_{k=0}^{\infty} \frac{1}{16^k}\left[\frac{4}{8k+1} - \frac{2}{8k+4} - \frac{1}{8k+5} - \frac{1}{8k+6}\right]$$



Du modèle à sa réalisation. La planche de Galton réalise-t-elle vraiment une distribution binomiale ?[1]

Bernard PARZYSZ

La planche de Galton est présentée classiquement comme un dispositif matériel permettant de simuler de façon manuelle une distribution binomiale et de faire apparaître visuellement une distribution de fréquences (du type diagramme en bâtons) voisine de la distribution de probabilité de cette loi. Cette utilisation s'appuie, plus ou moins consciemment, sur l'analogie existant entre les symétries de la planche à clous et celles d'un arbre probabilisé qui se ramifie par dichotomie. Cependant, une étude plus fine du dispositif fait apparaître que la situation est peut-être moins idyllique qu'elle n'y paraît. C'est une telle étude qu'a entreprise il y a quelques années Heinz STEINBRING, dans le but de présenter au niveau du lycée une définition *positive* de la notion de dépendance stochastique, laquelle est habituellement définie *négativement* par rapport à celle d'indépendance : deux événements A et B sont dits stochastiquement dépendants s'ils ne sont pas indépendants[2]. STEINBRING a proposé de substituer à cette approche *négative* de la dépendance une approche *positive*, basée sur une dialectique entre intuition et définition, au moyen précisément de la planche de Galton [STEINBRING 1986]. Mais cette étude peut également avoir d'autres utilités au niveau du lycée, entre autres celle de justifier, grâce à l'analogie formelle avec l'arbre probabilisé, certaines bases axiomatiques des probabilités et celle de pouvoir conduire les élèves à une démarche de recherche, dans le but de simuler *au mieux* le fonctionnement d'une planche réelle à l'aide d'un ordinateur.

I - Rappel et présentation didactique du modèle binomial classique

La loi binomiale est classiquement définie à partir de répétitions indépendantes d'une même expérience aléatoire conduisant à deux issues seulement

---

[1] Une partie de ce texte est reprise et adaptée de [ARTIGUE & PARZYSZ 2003].

[2] Remarquons que cette indépendance entre deux événements A et B, définie par $P(A \cap B) = P(A) \times P(B)$, est en réalité un *ménage à trois*, puisque qu'elle dépend de la probabilité P considérée. De plus, elle apparaît peu robuste : en effet, une variation, même minime, de $P(A)$ ou de $P(B)$ la fait disparaître. Ainsi, pour peu que $P(A)$ ou $P(B)$ ne soient connues qu'avec une légère incertitude, il devient impossible d'affirmer que les événements A et B sont indépendants.

(conventionnellement dénommées *succès* et *échec*). Le modèle de référence est celui de l'*urne bicolore* (ou *urne de Bernoulli*), c'est-à-dire une urne contenant N boules *indiscernables*, dont N_1 sont (par exemple) rouges et N_2 sont blanches ($N_1 + N_2 = N$) ; l'expérience aléatoire répétée consiste à tirer une boule de l'urne, *au hasard*, en la remettant ensuite dans l'urne (tirage dit *avec remise*, ou *non exhaustif*) et en mélangeant soigneusement les boules avant un nouveau tirage. Ceci pour assurer l'indépendance des succès ou échecs successifs, condition *sine qua non* pour modéliser à bon escient cette expérience par une loi binomiale.

N.B. : Les termes *au hasard*, *avec remise* et *non exhaustif* sont des conventions usuelles de vocabulaire destinées à indiquer que, dans ce modèle, toutes les boules de l'urne ont la même probabilité d'être tirées [PARZYSZ 1980]. Cette hypothèse découle en fait du *principe de raison insuffisante* de LEIBNIZ[3].

On effectue n tirages successifs ($n \geq 1$) et on s'intéresse au nombre de boules rouges tirées. La loi binomiale est définie par l'ensemble des probabilités d'obtenir k boules rouges, pour chaque entier k compris entre 0 et n.

Ce modèle est si classique qu'il n'est sans doute nul besoin de le développer ici du point de vue probabiliste, mais je voudrais néanmoins profiter de l'occasion pour en esquisser une présentation didactique possible, destinée à mettre en évidence les bases axiomatiques des probabilités[4]. Le premier ressort de cette présentation est l'arbre probabilisé (voir par exemple [PARZYSZ 1993], [GRANGÉ 2003]), qui présente physiquement une similitude forte avec les trajets possibles des boules sur une planche de Galton, similitude sur laquelle on peut s'appuyer pour étudier expérimentalement cette loi ; le second est le principe de symétrie, lequel se fonde sur l'*indiscernabilité* des boules, et donc sur les symétries de l'arbre (et qui est postulé).

a) Premier tirage

Considérons l'urne au départ (c'est-à-dire avant le premier tirage). Les N issues possibles (correspondant aux N boules contenues dans l'urne) peuvent être représentées par l'arbre ci-contre *(fig. 1)*.

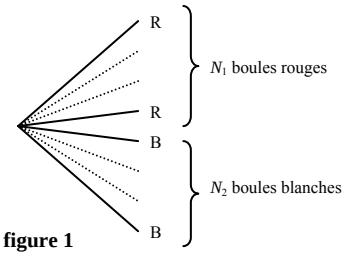


figure 1

En outre, le principe de symétrie (correspondant au fait que les boules sont *indiscernables*, c'est-à-dire qu'elles ont toutes la même probabilité d'être tirées)

[3] Repris ensuite par LAPLACE dans le but de justifier sa définition de la probabilité.

[4] Cette démarche est fondamentalement la même que celle utilisée par LAPLACE dans le second volume de sa *Théorie analytique des probabilités* [LAPLACE 1812].

implique que ces N issues sont équiprobables. En vertu du principe d'additivité[5], la probabilité d'obtenir une boule rouge (resp. blanche) lors du premier tirage est :

$$p = \frac{N_1}{N} \left(\text{resp. } q = \frac{N_2}{N} \right).$$

b) Second tirage

Une fois la première boule tirée et remise, effectuons un second tirage d'une boule (on est donc placé strictement dans les *mêmes conditions* que lors du premier tirage). La succession de ces deux tirages pourra être représentée par un arbre obtenu en concaténant, à chacune des N extrémités de l'arbre de la fig. 1, un arbre identique *(fig. 2)*.

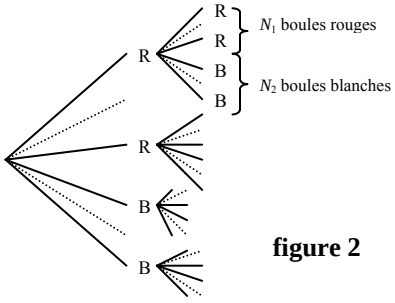


figure 2

Le nouvel arbre obtenu comporte ainsi $N \times N$ chemins menant de la racine à une extrémité. Toujours en vertu du principe de symétrie, on considère que tous ces chemins sont équiprobables[6]. Parmi ces chemins :

- $N_1 \times N_1$ correspondent au tirage de deux boules rouges (chemins RR),
- $N_1 \times N_2$ correspondent au tirage d'une boule rouge suivie d'une blanche (chemins RB),
- $N_2 \times N_1$ correspondent au tirage d'une boule blanche suivie d'une rouge (chemins BR),
- $N_2 \times N_2$ correspondent au tirage de deux boules blanches (chemins BB).

Il en résulte que la probabilité d'obtenir :

- deux boules rouges (chemin RR) est égale à $\frac{N_1 \times N_1}{N \times N}$, soit p^2 ;
- zéro boule rouge (chemin BB) est égale à $\frac{N_2 \times N_2}{N \times N}$, soit q^2 ;
- une boule rouge (chemins RB et BR) est égale à $2 \times \frac{N_1 \times N_2}{N \times N}$, soit $2pq$.

[5] Ce principe, aussi intuitif que celui de l'additivité des aires en géométrie, stipule que « la probabilité d'un événement donné est la somme des probabilités des événements élémentaires qui le constituent » : c'est la définition donnée en deuxième principe par LAPLACE (et qui figurait dans les programmes de première des années 1990).

[6] Hypothèse fondamentale et postulat de base en probabilités.

N.B.1 : Pour faciliter la représentation *physique* de la succession des tirages, on peut *condenser* l'arbre initial, ainsi que les suivants, en un arbre binaire, chacune des deux branches étant pondérée par le nombre de branches de l'arbre initial qu'elle représente [PARZYSZ 1993] *(fig. 3)*.

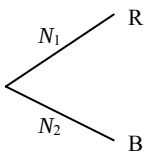


figure 3

L'arbre *condensé* correspondant à celui de la fig. 2 sera alors celui de la fig. 4 :

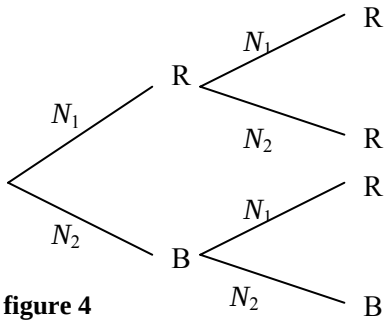


figure 4

N.B.2 : Nous venons ainsi de rencontrer un cas particulier de la *règle du produit* ($P(A \cap B) = P(A) \times P(B)$ quand les événements A et B sont indépendants) et de la justifier par la multiplication des chemins à partir de chaque nœud (suivie de la division par le nombre total de chemins possibles). Cette règle se généralise immédiatement, selon le même principe, au cas où l'expérience aléatoire conduit à plus de deux issues possibles (toutes équiprobables) et au cas où elle n'est pas répétée dans les mêmes conditions (tirage sans remise, par exemple), ce qui conduira à la formule $P(A \cap B) = P(A) \times P_A(B)$ [PARZYSZ 1993, pp. 100-101]. Il restera enfin à généraliser au cas où toutes les issues de l'expérience ne sont pas équiprobables (cas d'un dé pipé, par exemple), qui pourra être admis sur la foi des exemples précédents[7]. Mais revenons à nos moutons…

c) Tirages suivants

Plus généralement, imaginons maintenant une succession de n tirages. On pourra la représenter (au moins mentalement) comme précédemment, par concaténation d'arbres identiques à celui de la fig. 1. Cet arbre comportera N^n chemins (ou séquences) conduisant de la racine à l'une des extrémités, tous équiprobables en vertu du principe de symétrie. Ces chemins seront de la forme $A_1A_2\ldots A_n$, où chacun des A_k est égal à B ou à R. Rappelons en outre qu'à chaque

[7] De la même façon que, en s'appuyant sur la convergence des fréquences vers la limite théorique dans des cas simples, on admet l'existence d'une telle limite dans les autres cas (du type lancer de punaise).

tirage les boules rouges correspondent à N_1 branches possibles et les boules blanches à N_2 branches possibles *(fig. 1)*.

Chaque chemin représente donc une suite de n tirages comportant k boules rouges et $n-k$ boules blanches (k étant un entier compris entre 0 et n). Pour caractériser un tel chemin, il faut :

- fixer, parmi les n tirages effectués, les k tirages qui donneront les k boules rouges (il y en a $\binom{n}{k}$, nous dit la combinatoire)
- puis choisir, pour chacun des n tirages, une boule parmi celles qui sont de la *bonne* couleur (il y en a donc N_1 ou N_2, selon le cas), ce qui nous donne au total $N_1^{\,k} \times N_2^{\,n-k}$ chemins. D'où en définitive $\binom{n}{k} N_1^{\,k} \times N_2^{\,n-k}$ chemins présentant k boules rouges (sur un total de N^n) et – toujours en vertu du principe de symétrie – la probabilité d'obtenir k boules rouges en n tirages est égale à $\binom{n}{k} \frac{N_1^{\,k} \times N_2^{\,n-k}}{N^n}$, soit $\binom{n}{k} p^k\, q^{n-k}$.

Refermons cette parenthèse didactique et formalisons le résultat précédent : soit la variable aléatoire X égale au nombre total de boules rouges obtenues au cours des n tirages. Quelque soit l'entier k compris entre 0 et n, l'événement $\{X = k\}$ a pour probabilité $\mathrm{P}(X = k) = \binom{n}{k} p^k\,(1-p)^{n-k}$.

La loi de la variable X est appelée *loi binomiale* de paramètres n et p, et est notée $\mathcal{B}(n\,,p)$.

II - De l'arbre à la planche : le modèle théorique

La *planche de Galton*[8] est un dispositif matériel classique, dont le but est de mettre en évidence la distribution binomiale[9] dans le cas $p = 0{,}5$. Il consiste en une planche inclinée sur laquelle des clous sont plantés en quinconce *(fig. 5)*. En haut et au centre de la planche sont lâchées des billes, qui rebondissent sur les clous avant d'aboutir dans des godets alignés en bas.

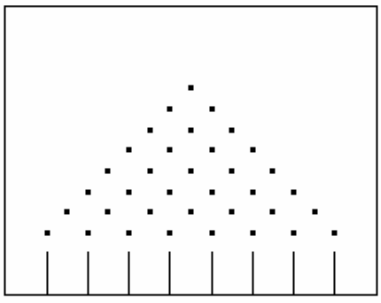


figure 5

[8] Sir Francis GALTON (1822-1911), voyageur et physiologiste anglais, s'intéressa également à l'anthropologie, à la météorologie… et à la statistique.

[9] Et, au-delà, sa convergence vers la loi normale.

En tenant compte des symétries (locales et globales) que présente la planche - et qui sont en quelque sorte analogues à celles de l'arbre non condensé du paragraphe précédent –, le modèle théorique habituellement donné de ce dispositif est le suivant : la bille lâchée heurte le clou du haut (rang 1) et rebondit vers l'un des deux clous du dessous (rang 2), avec une probabilité égale de se diriger vers l'une ou vers l'autre (disons, pour simplifier : vers la droite ou vers la gauche). Le même processus se répète à chaque rang jusqu'au dernier (rang n), après quoi la bille tombe dans un des $n+1$ godets placés en dessous. La bille a ainsi été soumise à n répétitions d'une même expérience aléatoire comportant deux issues (aller vers la droite ou vers la gauche) auxquelles, dans un premier temps (nous y reviendrons), on attribue la même probabilité. On suppose aussi que d'un clou à l'autre, les rebonds à droite ou à gauche sont indépendants.

Ce modèle – pensons à l'arbre – est alors analogue au modèle d'urne de Bernoulli avec remise décrit plus haut, *dans le cas où il y a autant de billes de chaque couleur*. En effet, l'arrivée de la bille sur un clou du rang k est comparable avec le k-ième tirage dans l'urne :

a) à chaque fois, l'expérience est répétée dans les mêmes conditions ;

b) il y a deux issues possibles, et deux seulement : la bille se dirige vers la droite ou vers la gauche ;

c) chacune de ces deux issues est affectée de la même probabilité ;

d) on peut convenir que l'événement « la bille se dirige vers la gauche (resp. vers la droite) » (planche de Galton) correspond à l'événement « la boule tirée est rouge (resp. blanche) » (urne bicolore).

Numérotons les godets de 0 à n de la gauche vers la droite (ou inversement) et considérons la variable aléatoire Y égale au numéro du godet dans lequel tombe la bille (c'est-à-dire aussi au nombre de rebonds vers la droite effectués par la bille au cours de son trajet). La probabilité que la bille tombe dans le godet k est donc $\mathrm{P}(Y=k)=\binom{n}{k}\left(\frac{1}{2}\right)^k\left(\frac{1}{2}\right)^{n-k}$. Dans l'hypothèse d'équiprobabilité qui a été faite, on peut par conséquent dire que la variable Y suit la loi binomiale $\mathcal{B}\left(n,\frac{1}{2}\right)$.

III - La planche de Galton : du modèle théorique à sa réalisation pratique

Il se trouve cependant que, dans une réalisation concrète de l'expérience, ce n'est pas exactement cette répartition que l'on observe. Voici un exemple obtenu

par STEINBRING avec 3 000 billes et une planche à 8 rangs de clous (donc 9 godets)[10] :

Godet	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Distribution observée	8	74	296	712	883	667	283	69	8
Distribution binomiale	11,7	93,7	328,1	656,2	820,3	656,2	328,1	93,7	11,7

Au risque de 5 %, les deux distributions sont statistiquement différentes ($\chi^2 = 32{,}05$ alors que la valeur limite est 15,51[11]). On s'aperçoit qu'il y a en fait une plus forte concentration de billes vers le centre que ce qu'indiquerait la loi binomiale du modèle théorique (au détriment des bords).

Une analyse *physico-géométrique* de la planche de Galton fait penser que les chemins en zigzag risquent d'être un peu plus fréquents que les chemins en ligne droite : en effet, si la bille arrive de la droite sur un clou, elle percutera ce clou du côté droit et rebondira ensuite plus fréquemment vers la droite, et inversement lorsqu'elle arrive de la gauche *(fig. 6)*.

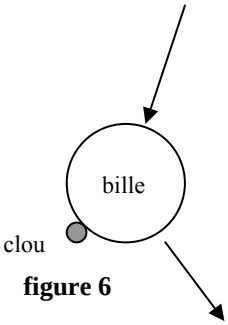


figure 6

Ceci remet en cause l'indépendance des différents niveaux de la planche : en effet, les considérations qui précèdent permettent au contraire de penser qu'il existe une (légère) dépendance dans la succession des rencontres de la bille avec les clous, liée aux caractéristiques physiques du dispositif : rapport des diamètres des clous et des billes, écartement des clous, écartement des rangées de clous, inclinaison de la planche, etc.

Nous utiliserons les notations suivantes (le schéma de la figure 7 permet de visualiser la situation) :

a) $P(n\,,k)$ (resp. $Q(n\,,k)$) est la probabilité que la bille arrive sur le $k^{\text{ème}}$ clou de la $n^{\text{ème}}$ rangée (clou noté $[n\,,k]$) sachant qu'à la $(n-1)^{\text{ème}}$ rangée, elle avait heurté le clou $[n-1\,,k]$ (resp. $[n-1\,,k-1]$).

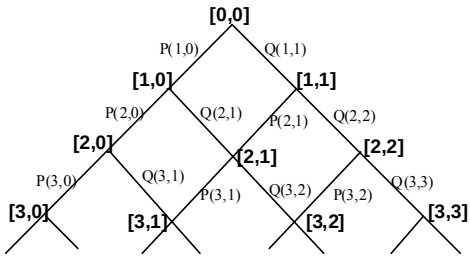


figure 7

b) On pose $p = 0{,}5 - \varepsilon$ et $q = 0{,}5 + \varepsilon$.

[10] in [STEINBRING 1986, p. 27]

[11] Voir l'article de BONNEVAL, L. M., HENRY, M. : *Tests d'adéquation à une loi équirépartie et autres tests de Khi-deux* dans ce même ouvrage.

On peut alors imaginer un modèle [HENNEQUIN 1981] conduisant à un système récurrent du type :

$$\begin{cases} P(1\,;0) = Q(1\,;1) = 0{,}5 \\ P(n\,;n) = Q(n\,;0) = 0 \ \text{ pour tout } n \\ P(n\,,k) = p\,P(n-1\,,k) + q\,Q(n-1\,,k) \ \text{ pour } \ 0 \le k \le n-1 \\ Q(n\,,k) = q\,P(n-1\,,k-1) + p\,Q(n-1\,,k-1) \ \text{ pour } \ 1 \le k \le n \end{cases}$$

Par approximations successives, STEINBRING aboutit au résultat suivant, en prenant $\varepsilon = 0{,}037$:

Godet	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Distribution observée	8	74	296	712	883	667	283	69	8
Distribution binomiale	6,8	71	301,6	678,8	883	678,8	301,6	71	6,8

Comme on peut le constater, la distribution théorique est, dans ce cas, nettement plus voisine de la distribution observée que la distribution binomiale[12].

STEINBRING fait observer que, dans cette démarche, la dépendance n'est pas introduite de façon gratuite, mais résulte d'une analyse des conditions concrètes de l'expérience ; la fonction du concept d'indépendance stochastique est ici de contrôler l'adéquation du modèle théorique binomial au résultat expérimental, selon le schéma de la *fig. 8* :

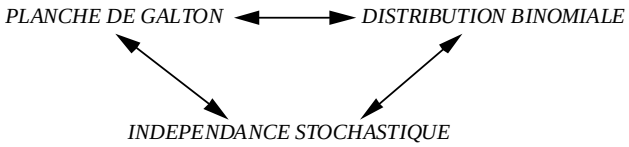


figure 8

N.B. : De façon analogue, en étudiant les travaux de GEIßLER (1889) qui a enregistré 4 millions de naissances[13], on a constaté que, parmi les familles de 12 enfants, il y a davantage de familles comportant beaucoup d'enfants du même sexe que ce qu'on s'attendrait à trouver d'après le modèle binomial classique $B(N\,,0{,}5168)$[14], et qu'en outre la distribution observée est un peu asymétrique. Dans ce cas également, l'hypothèse d'une légère dépendance dans

[12] Un test du χ^2 le confirme : on trouve cette fois $\chi^2 = 3{,}69$, pour une valeur limite de 15,51 au risque de 5 %.

[13] Arthur GEIßLER, docteur en médecine, publia en 1889, dans la revue du Bureau de la Statistique de Saxe, un article dans lequel il donne des statistiques sur les naissances en Saxe (nombre de garçons et de filles, âge de la mère, etc.), établie sur la période 1876-1885 (renseignement communiqué par H. STEINBRING).

[14] On prend habituellement 0,5168 comme fréquence de naissance des garçons (ils sont plus nombreux au départ, mais résistent moins bien au temps…).

la suite des naissances d'une même famille permet de *coller* davantage aux observations (mais une hypothèse génétique *explicative* fait encore malheureusement défaut).

Cette étude avait en fait trois finalités principales. La première était d'étudier de façon un peu fine la distribution statistique des billes donnée par une planche de Galton, dont il est couramment admis, sur la foi de considérations de symétrie liées à la structure d'un dispositif idéal, qu'elle réalise une bonne approximation de la loi binomiale dans le cas $p = 0{,}5$. Nous avons vu en fait, d'une part que certaines réalisations physiques peuvent s'en écarter notablement, et d'autre part qu'il est possible de trouver à la fois une explication à l'écart observé et un modèle plus convenable. En effet, l'ordinateur permet de simuler aisément le fonctionnement d'une planche *idéale* ; il fournit alors une réalisation du modèle binomial, ce qui n'est guère étonnant puisque c'est en fait ce modèle qui a été introduit initialement dans la machine.

La seconde finalité était d'aborder – via les arbres probabilisés – la question de la justification de la règle du produit, premier pas vers la formule des probabilités composées. Enfin, la troisième finalité était – à la suite de STEINBRING – d'indiquer une approche expérimentale *positive* de la notion de dépendance stochastique : pour deux événements liés à une même expérience aléatoire, c'est en effet la dépendance qui est la règle et l'indépendance l'exception. La planche de Galton en fournit un excellent exemple, puisque – lorsqu'on y réfléchit – la façon dont la bille rebondit *dépend* (au sens courant) de la façon dont elle vient frapper le clou, et donc du clou précédemment heurté.

Pour conclure, je ne résisterai pas au plaisir de laisser le mot de la fin à D'ALEMBERT[15] : *« Cela est digne, ce me semble, de l'attention des calculateurs, et irait à réformer bien des règles unanimement reçues sur les jeux de hasard ».* Cela me semble également digne de l'attention des enseignants de terminale, qui pourraient faire utiliser la planche de Galton en classe pour permettre à leurs élèves d'étudier expérimentalement un dispositif aléatoire non trivial et d'en élaborer des modèles possibles, lesquels pourraient ensuite être simulés à l'aide de l'ordinateur dans le but de les tester[16].

---

[15] Jean Le Rond d'ALEMBERT (1717-1783), mathématicien et philosophe, principal auteur avec Denis DIDEROT de la *Grande Encyclopédie*. La citation est extraite de l'article *Croix ou Pile* de ce monumental ouvrage.

[16] On pourrait par exemple, à l'instar de H. STEINBRING, faire fonctionner réellement une planche de Galton, puis réfléchir sur le phénomène de rebond avant de simuler cette expérience à l'aide d'un ordinateur en introduisant un p dans les probabilités de rebond, puis chercher à ajuster ce p pour obtenir une simulation plus satisfaisante (c'est-à-dire plus *conforme* à l'expérience physique).

Bibliographie

ARTIGUE, M. & PARZYSZ, B. (2003) : Causalités et dépendances : quelle place dans les recherches en didactique des mathématiques, in Viennot (éd.) : *Enquête sur le concept de causalité*. Coll. Science, Histoire et Société. Presses Universitaires de France, pp. 123-151.

GRANGÉ, J.-P., (2003) : Arbres et tableaux en probabilités conditionnelles, in *Probabilités au lycée*, APMEP, pp. 91-124.

HENNEQUIN, P.-L., (1981) : Schéma de Bernoulli et planchettes à clous, in *Bulletin de l'APMEP*, pp. 435-441.

LAPLACE, P.-S., (1812) : *Théorie analytique des probabilités*, Tome VII des *Œuvres de Laplace*. Paris, Imprimerie royale (1847), pp. 195 sq, Réédition Jacques Gabay, 1995.

PARZYSZ, B., (1980) : Les mots pour le dire. Sur le vocabulaire des dénombrements, in *Groupe français-mathématiques. Volume 2*, IREM Université Paris 7, pp. 133-38.

PARZYSZ B., (1993) : Des statistiques aux probabilités : exploitons les arbres, in *Repères-IREM* n° 10, pp. 91-104.

STEINBRING, H., (1986) : L'indépendance stochastique. Un exemple de renversement du contenu intuitif d'un concept et de sa définition mathématique formelle, in *Recherches en Didactique des Mathématiques* vol. 7 n° 3, pp. 5-50.

Phénomènes gaussiens et lois normales[1]

Michel HENRY

I - Phénomènes gaussiens, origine de la loi normale

La loi normale est la loi de certains phénomènes continus qui fluctuent autour d'une valeur moyenne μ, de manière aléatoire, résultante d'un grand nombre de causes additives et indépendantes. C'est une illustration du théorème le plus important du calcul des probabilités, le théorème-limite central (TLC)[2].

La dispersion des valeurs observées d'un caractère gaussien est représentée par un écart-type σ. L'observation statistique de ce type de phénomènes conduit à la remarque que 95 % environ des observations se situent dans l'intervalle $]\mu - 2\sigma ; \mu + 2\sigma[$, dit *plage de normalité*. (Doc. GEPS d'accompagnement du programme de 1ère L).

La loi normale est, entre autres, la loi des erreurs de mesure d'un phénomène physique. Historiquement, c'est par l'étude du comportement de ces erreurs que LAPLACE (1774, 1809) puis GAUSS (1809) ont mis en évidence l'expression mathématique de cette loi[3].

Mais le précurseur fut Abraham DE MOIVRE (1733). Il cherchait à calculer avec précision les probabilités binomiales pour améliorer le contrôle de l'écart $|F_n - p|$ entre fréquence observée et probabilité limite dans le problème de Bernoulli (cf. le § IV ci-dessous). DE MOIVRE appliqua sa formule complétée par son ami STIRLING donnant un équivalent de n !, et obtint l'expression en $\exp(-x^2)$ comme méthode de calcul pour une meilleure approximation numérique de ces probabilités. (Théorème de Moivre-Laplace).

II - Modèle probabiliste de la loi normale

Pour modéliser un phénomène gaussien, nous allons distinguer sa description heuristique du modèle probabiliste que l'on peut proposer pour le représenter.

---

[1] Ce chapitre est repris d'un article paru dans *Repères-IREM* n° 51 d'avril 2003.

[2] Dans l'article suivant, *Théorie des erreurs, courbe en cloche et normalité*, Jean-François PICHARD dévoile les contextes historiques dans lesquels les dénominations de *loi normale*, de *Laplace-Gauss* ou du *théorème-limite central* ont été attribuées.

[3] Pour une étude détaillée de la genèse historique de la théorie des erreurs, voir l'article suivant.

Hypothèses de travail

On étudie un caractère aléatoire continu (en dimension 1), pouvant prendre n'importe quelle valeur réelle, symétriquement autour d'une valeur centrale μ. Une étude statistique montre que, pour un grand nombre d'observations, 95 % d'entre elles se situent dans un intervalle centré en μ et de demi-longueur 2σ.

Pratiquement, dans beaucoup d'applications, les valeurs possibles du caractère sont nécessairement positives (mesures de grandeurs, quantités…). Mais il n'y a pas de gros inconvénient à proposer un modèle continu sur tout $\mathbb{R}$, si la probabilité est surtout concentrée autour d'une moyenne μ positive, car dans ce modèle gaussien la probabilité de l'intervalle $]-\infty\,;0[$ est négligeable. Cette propriété du modèle rendra compte du fait que la quasi-totalité des valeurs observées sont concentrées autour de cette valeur centrale.

Hypothèses de modèle

On prend donc comme ensemble référentiel, $\Omega = \mathbb{R}$. Soit X une variable aléatoire définie sur Ω à valeurs dans $\mathbb{R}$ représentant les valeurs possibles du caractère étudié.

X suit une loi normale, si sa densité f_X est donnée par :

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \text{ pour tout x réel.}$$

Rappelons que la loi P_X d'une variable continue X, de densité f_X est entièrement déterminée par les probabilités

$$P_X(]a\,,b[) = P(X \in]a\,,b[) = \int_a^b f_X(x)\,dx.$$

Contrairement aux cas des lois uniformes ou exponentielles, la densité de la loi normale est l'outil incontournable pour déterminer les probabilités liées à la variable normale X, car sa fonction de répartition, définie par :

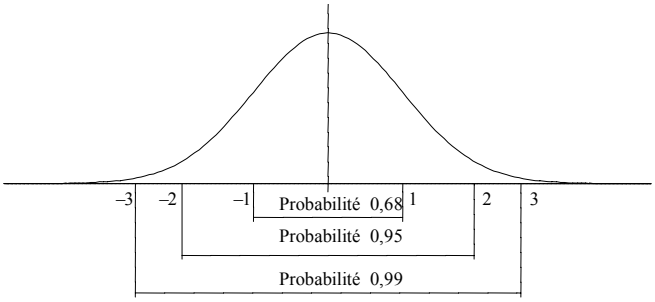
$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(x)\,dx.$$

ne peut être exprimée par les fonctions élémentaires.

Cette loi est désignée par $\mathcal{N}(\mu\,,\sigma)$. On ramène toutes les lois normales à un même type de base, $\mathcal{N}(0\,,1)$, en *centrant et réduisant* la variable X : on pose $U = \frac{X-\mu}{\sigma}$. La densité f_U de la loi de U est alors définie par : $f_U(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$ pour tout u réel. Sa fonction de répartition est notée Φ :

$$\Phi(u) = P(U \leq u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\,dt.$$

La représentation graphique de f_U est la fameuse courbe en cloche dite *courbe de Gauss* :



On a de manière précise $P(|U| \leq 1{,}65) = 0{,}9$ et $P(|U| \leq 1{,}96) = 0{,}95$.

III - Le théorème-limite central

Nous allons voir pourquoi cette loi s'impose *naturellement*. Anticipant quelque peu sur l'article suivant, reprenons la question historique des erreurs de mesure. Mesurer une grandeur consiste à mettre en œuvre un phénomène physique qu'un instrument de mesure est susceptible d'enregistrer sous la forme d'une donnée numérique, dont la précision dépend de la sensibilité de l'appareil et des aléas qui accompagnent inéluctablement toute procédure de mesure. Soit X_0 la variable aléatoire qui, à une telle opération, fait correspondre la valeur affichée par l'appareil.

Remarque de physicien : si on recommence n fois la même mesure, on obtient n valeurs entachées des *erreurs de mesure*. En prenant leur moyenne, on obtient une valeur plus précise, ayant amélioré virtuellement la sensibilité des appareils en réduisant la dispersion des résultats. Cette propriété, connue depuis longtemps, peut être expliquée simplement par un petit raisonnement probabiliste.

En effet, si on désigne par X_i les différentes variables aléatoires prenant pour valeurs les mesures successives, ces X_i sont des répliques indépendantes de la variable aléatoire X_0 et constituent un échantillon $X = (X_1 , X_2 , \dots , X_n)$. La valeur retenue pour la grandeur à mesurer est alors la moyenne des X_i :

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

On peut faire deux remarques :

- Si on suppose que les X_i se répartissent aléatoirement autour de la valeur μ à mesurer (on suppose qu'il n'y a pas d'erreur systématique : $E(X_i) = \mu$), il en est de même pour $\overline{X}$ dont la valeur moyenne est :

$$E\left(\overline{X}\right) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{n\,\mu}{n} = \mu.$$

– Mais, tenant compte de l'indépendance des X_i, on a, d'après une propriété de base de la variance,

$$\mathrm{Var}\left(\overline{X}\right) = \mathrm{E}\left[\left(\overline{X} - \mathrm{E}\left(\overline{X}\right)\right)^2\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{n\,\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

La dispersion de $\overline{X}$ est donc représentée par l'écart-type $\frac{\sigma}{\sqrt{n}}$.

$\overline{X}$ est donc plus concentrée autour de μ que chacune des X_i, et ceci d'autant plus que n est grand : avec $n = 100$, on gagne un ordre de grandeur sur la sensibilité des instruments. Un théorème (facile) dit que si les X_i sont des variables normales de loi $\mathcal{N}(\mu\,,\,\sigma)$, alors $\overline{X}$ est aussi une variable normale, mais de loi $\mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

Le **théorème-limite central (TLC)**, théorème beaucoup plus fort, dit que[4] :

Si $(X_n)_{n\in\mathbb{N}}$ *est une suite de variables aléatoires indépendantes et de même loi, telles que* $\mathrm{E}(X_n) = \mu$ *et* $\mathrm{Var}(X_n) = \sigma^2$*, alors la suite des variables moyennes réduites* $Z_n = \frac{\overline{X_n} - \mu}{\sigma}\sqrt{n}$ *converge en loi vers une variable normale* U *centrée réduite :* $U \sim \mathcal{N}(0\,,\,1)$.

La loi commune des X_n peut être tout à fait quelconque, la seule condition est que son espérance et sa variance existent. Pratiquement, on peut interpréter cet énoncé par la propriété de convergence suivante : Pour tout intervalle $]a, b[$, $\mathrm{P}(Z_n \in]a\,,\,b[) \xrightarrow[n\to\infty]{} \mathrm{P}(U \in]a\,,\,b[) = \Phi(b) - \Phi(a) = \int_a^b \frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}u^2}\,du.$

Même si on ne sait rien sur le comportement des X_i, le TLC dit que pour n assez grand (en général, on admet avoir une assez bonne précision dès que $n > 50$), on connaît approximativement la loi de $\overline{X}$: c'est presque une loi normale $\mathcal{N}(\mu\,,\,\frac{\sigma}{\sqrt{n}})$.

En particulier, si $\mu > 0$, $\mathrm{P}(\overline{X_n} < 0)$ est approximé par $\Phi\left(-\frac{\mu}{\sigma}\sqrt{n}\right)$ qui tend vers 0 quand n tend vers l'infini ; ceci justifie notre remarque antérieure sur la possibilité d'utiliser des approximations normales même pour des lois intrinsèquement positives, dès lors que les valeurs de μ et σ permettent de négliger la probabilité de l'événement $\left\{\overline{X_n} < 0\right\}$.

---

[4] Dans cet énoncé du TLC, $\overline{X_n}$ désigne la moyenne arithmétique des n premières variables X_i, et *converge en loi* veut dire que la suite des fonctions de répartition des Z_n converge simplement vers celle de U : pour tout z réel, $\mathrm{P}(Z_n \le z) \to \mathrm{P}(U \le z)$ quand $n \to \infty$.

Le TLC est un théorème fondamental en statistique inférentielle. Les valeurs observées sur un n-échantillon (aléatoire à prélèvements indépendants) sont considérées (modélisation) comme des valeurs prises par n variables aléatoires X_i indépendantes et de même loi qu'une variable parente X, en général inconnue. Dans de nombreuses applications, la moyenne des X_i intervient et, si n est assez grand ($n > 50$ au minimum), la loi de cette moyenne est, d'après le TLC, suffisamment proche d'une loi normale $\mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ pour la précision attendue dans les applications standard.

En général, μ est inconnue (c'est sa valeur qui fait l'objet du problème statistique posé), mais σ est connu avec assez de précision. On peut alors calculer avec une approximation suffisante des probabilités (qui ne dépendent pas de μ) d'événements liés à la moyenne centrée $\overline{X} - \mu$, dont la loi est approchée par une loi normale $\mathcal{N}(0\,,\sigma)$. Ce calcul permet notamment :

- de déterminer un intervalle de confiance pour μ à un niveau de confiance $1 - \alpha$ donné, c'est-à-dire de trouver un $\varepsilon > 0$ tel que :

  $$P(\overline{X} \in]\,\mu - \varepsilon, \mu + \varepsilon[) \geq 1 - \alpha.$$

  Cette probabilité étant proche de $P(|U| < \frac{\varepsilon}{\sigma})$, on peut prendre pour $\frac{\varepsilon}{\sigma}$ le quantile correspondant de la loi normale centrée réduite.

  En prenant un *risque* inférieur à α de se tromper, on peut alors affirmer que $\mu \in]\,\overline{X} - \varepsilon\,,\,\overline{X} + \varepsilon[$.

- ou de faire un test d'hypothèse, désirant tester une certaine hypothèse H_0 relative à la moyenne μ.[5]

  Par exemple, un μ_0 étant donné, on veut tester l'hypothèse $H_0 : \mu = \mu_0$ contre l'hypothèse $H_1 : \mu > \mu_0$. On convient que la décision de rejet de H_0 interviendra lorsque la moyenne $\overline{x}$ des valeurs observée sur l'échantillon sera supérieure à une *valeur critique* $\mu_c > \mu_0$. La valeur critique μ_c est déterminée par un *seuil* α donné, de telle sorte que, sous l'hypothèse H_0, la probabilité que $\overline{X}$ dépasse μ_c est inférieure à α.

  Cette condition de confiance s'écrivant $P_{\mu_0}(\overline{X} - \mu_0 > \mu_c - \mu_0) \leq \alpha$, le TLC permet de calculer μ_c tel que $P\left(U > \frac{\mu_c - \mu_0}{\sigma}\right) \leq \alpha$. Ainsi, avec une moyenne $\overline{x}$ des valeurs observées dans l'échantillon supérieure à μ_c, on rejette

[5] Pour une présentation plus complète des tests d'hypothèses, on pourra se reporter à l'article *Introduction aux tests d'hypothèses et exemples* de Michel HENRY et Annette CORPART dans ce même volume.

l'hypothèse H_0 pour décider que $\mu > \mu_0$ au seuil de signification α (i.e. en prenant un risque inférieur à α de se tromper).

Le TLC est donc un outil puissant de contrôle de phénomènes qui sont des moyennes arithmétiques. La loi normale s'impose dans son énoncé[6]. Sa démonstration fait intervenir des outils mathématiques de haut niveau accompagnant les lois continues (transformation de Fourier).

IV - Exemple historique de base : le théorème de Moivre - Laplace

Plaçons nous dans la situation du problème de Bernoulli : on répète n fois une même épreuve de Bernoulli (expérience aléatoire à 2 issues, succès de probabilité p ou échec de probabilité $(1 - p)$).

On sait (théorème de Bernoulli) que les fréquences F_n des succès *tendent à se stabiliser* vers p quand n devient très grand. Plus précisément, $P(|F_n - p| < \varepsilon) \to 1$ quand $n \to \infty$.

Mais comment contrôler cette *stabilisation*, c'est-à-dire le degré de précision obtenu quand on évalue p par la valeur observée de la fréquence F_n ?

La technique consiste à construire un intervalle de confiance pour p, dont les bornes dépendent de la valeur observée de la fréquence des succès pour un n donné.

Modélisons cela. A chaque épreuve, on associe la variable aléatoire X_i de Bernoulli définie par :

- $X_i = 1$ si on a un succès (de probabilité p),
- $X_i = 0$ si on a un échec (de probabilité $(1 - p)$).

On a $E(X_i) = p$ et $Var(X_i) = p\,(1 - p)$.

Or, $F_n = \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$, puisque la somme est égale au nombre de succès obtenus en n épreuves. D'où, $E(F_n) = p$ et $Var(F_n) = \frac{p(1-p)}{n}$ (cf. le calcul indiqué au début de ce paragraphe).

On peut donc appliquer le TLC à F_n, ce qui donne le théorème de Moivre-Laplace que l'on peut *moderniser* dans cet énoncé condensé :

Si n est assez grand, la loi de F_n est proche de la loi normale $\mathcal{N}\left(p\,,\sqrt{\frac{p\,(1-p)}{n}}\right)$

[6] D'autres lois ont été tentées pour modéliser les erreurs comme le montre l'article suivant. La loi normale a fini par s'imposer.

Pratiquement, cette approximation est acceptable dès que $np(1-p) \geq 5$. (en général, on peut appliquer le TLC quand $n \geq 50$, en supposant ici que p ne soit pas trop voisin de 0 ou de 1, mais le contexte binomial permet cette condition plus précise). Comme $n F_n$ est égal au nombre de succès en n épreuves de Bernoulli, $n F_n$ est une variable binomiale de loi $\mathcal{B}(n, p)$. Le théorème de Moivre-Laplace donne l'approximation en loi de la loi binomiale : $\mathcal{B}(n, p) \approx \mathcal{N}\left(np, \sqrt{np(1-p)}\right)$.

On peut aussi énoncer cette propriété en considérant la fréquence réduite $Z_n = \frac{F_n - p}{\sqrt{p(1-p)}}\sqrt{n}$: la loi de Z_n est proche d'une loi normale centrée réduite.

V - Intervalle de confiance pour une proportion (sondages)

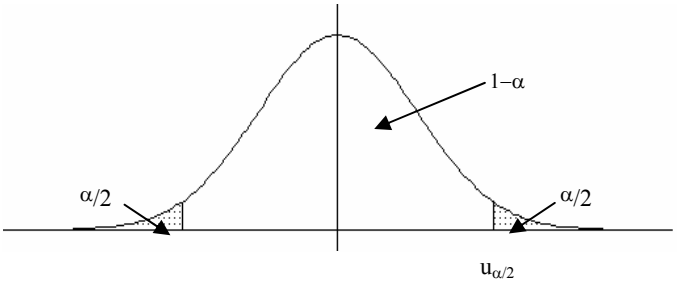
Ce qui précède fournit la démarche de base pour déterminer des intervalles de confiance pour des sondages aléatoires simples[7].

Soit p la proportion des éléments considérés dans une population. Pour estimer p, on en extrait au hasard un échantillon de taille n dans lequel la fréquence de ces éléments est F_n.

Un niveau de confiance $1-\alpha$ étant fixé (par exemple 0,95), il convient de déterminer un ε de telle sorte que la probabilité que p soit dans l'intervalle de confiance $]F_n - \varepsilon, F_n + \varepsilon[$ soit supérieure à $1-\alpha$. La précision ε de cette estimation est donc la plus petite valeur possible pour ε vérifiant la condition de confiance :

$$P(p \in]F_n - \varepsilon, F_n + \varepsilon[) = P\left[|Z_n| < \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right] \approx P\left[|U| < \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right] \geq 1-\alpha.$$

Cette condition montre que la valeur de $\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}$ peut être obtenue à partir du fractile $u_{\alpha/2}$ d'ordre $\frac{\alpha}{2}$ de la loi $\mathcal{N}(0, 1)$ de U défini par $P(U > u_{\alpha/2}) = \frac{\alpha}{2}$, ainsi que l'indique le schéma suivant.



[7] Pour de plus amples développements sur les sondages et leurs simulations, on pourra se reporter au volume 2 *Activités statistiques pour la classe*.

On a donc $P(|U| \leq u_{\alpha/2}) = 1 - \alpha$. Par exemple, pour $\alpha = 0{,}05$, on trouve $u_{\alpha/2} = 1{,}96$ (lecture dans la table).

La condition de confiance est donc réalisée avec $u_{\alpha/2} = \dfrac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}$, d'où $\varepsilon = \dfrac{u_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}$, variant comme $\dfrac{1}{\sqrt{n}}$.

Mais p est inconnu puisque c'est la proportion à estimer ! On peut cependant majorer $\sqrt{p(1-p)}$, car si on augmente ε, on élargit l'intervalle de confiance et la probabilité $P(p \in]F_n - \varepsilon\,,\, F_n + \varepsilon[)$ augmente, la condition de confiance est alors d'autant mieux vérifiée.

Or, pour tout $p \in]0\,,\,1[$, $p(1-p) \leq \dfrac{1}{4}$. Cette majoration est acceptable si p est assez voisin de $\dfrac{1}{2}$ (entre 0,3 et 0,7). De plus, si $\alpha = 0{,}05$, on a $u_{\alpha/2} = 1{,}96 < 2$, d'où $\varepsilon < \dfrac{1}{\sqrt{n}}$ est une majoration possible dans ces conditions. On obtient alors la formule simplifiée de l'intervalle de confiance donnant la fourchette proposée dans un thème d'études du programme de seconde :

$$P\left(p \in \left]F_n - \frac{1}{\sqrt{n}}\,;\, F_n + \frac{1}{\sqrt{n}}\right[\right) \geq 0{,}95$$

De manière plus générale[8], on a une forme plus précise de l'intervalle de confiance pour la proportion p au niveau $1 - \alpha$:

$$P\left(p \in \left]F_n - \frac{u_{\alpha/2}\sqrt{F_n(1-F_n)}}{\sqrt{n}}\,;\, F_n + \frac{u_{\alpha/2}\sqrt{F_n(1-F_n)}}{\sqrt{n}}\right[\right) \approx 1 - \alpha.$$

D'où la *fourchette de sondage* théorique pour estimer p au niveau de confiance $1 - \alpha$, à partir d'un échantillon de taille n pour lequel la valeur observée de F_n est f_n :

$$\left]f_n - \frac{u_{\alpha/2}\sqrt{f_n(1-f_n)}}{\sqrt{n}}\,;\, f_n + \frac{u_{\alpha/2}\sqrt{f_n(1-f_n)}}{\sqrt{n}}\right[.$$

---

[8] On peut montrer que $\dfrac{F_n - p}{\sqrt{F_n(1-F_n)}}\sqrt{n}$ converge aussi en loi vers une variable normale centrée réduite (application d'un lemme sur la convergence en loi, en remarquant que $\dfrac{F_n(1-F_n)}{p(1-p)} \xrightarrow{p.s.} 1$ d'après la loi forte des grands nombres). La précision de l'estimation en découle.

Théorie des erreurs, courbes en cloche et normalité

Jean-François PICHARD

Résumé : J'indique d'abord en hors-d'oeuvre quelques courbes géométriques en forme de cloche. Ensuite j'étudie brièvement l'évolution des idées concernant ce qui est appelé maintenant en statistique la théorie de base de l'estimation, partant du théorème de J. BERNOULLI pour aller à la méthode des moindres carrés et l'aboutissement des recherches en théorie des erreurs : le Théorème-Limite Central, en passant par BAYES et LAPLACE sur les probabilités *a posteriori*. Je mentionne enfin quelques recherches en sciences sociales et en biométrie au XIX^{e} siècle et début du XX^{e} siècle, centrées autour de la *loi normale*.

Introduction

La représentation statistique d'observations d'un phénomène, qui est faite actuellement dans l'enseignement secondaire, s'appuie essentiellement sur la loi normale et le graphe de sa densité, la fameuse *courbe en cloche*. Indiscutablement, le rôle de la loi normale est primordial en statistique ; cependant, toute courbe en forme de cloche n'est pas nécessairement le graphe de la densité d'une loi normale, comme nous en donnerons quelques exemples plus loin. La loi normale (nous revenons sur cette dénomination dans les remarques qui terminent cette introduction) est bien sûr pratiquement la distribution la plus importante dans les applications, mais il ne faudrait pas reproduire dans l'enseignement les errements qui ont eu cours pendant une grande partie du XIX^{e} siècle, où les chercheurs de diverses disciplines (sociologie, biométrie, psychologie,...), qui utilisaient les méthodes statistiques, cherchaient à ramener leurs analyses à la loi normale, comme cela se produisait dans les sciences d'observation.

L'invention de la loi normale et une caractérisation de son champ d'intervention sont le premier aboutissement d'une recherche entamée dans la première moitié du XVIII^{e} siècle concernant les erreurs d'observations et le meilleur milieu à prendre entre des observations discordantes. Cette théorie des erreurs a émergé des sciences les plus anciennes, la *science reine* : l'astronomie, et la géodésie dont les méthodes de mesure étaient alors les plus précises parmi les sciences d'observation.

Remarques sur la dénomination *loi normale*

Le nom de l*oi normale* a été donné et popularisé par Karl PEARSON en 1893 (Francis GALTON avait déjà parlé de *courbe normale* en 1889 dans *Natural Inheritance*) à la loi appelée par l'école française *deuxième loi de Laplace* ou *loi de Laplace-Gauss* et par l'école anglo-saxonne *loi de Gauss*, en raison de la difficulté d'attribution du premier découvreur : A. DE MOIVRE, P. S. LAPLACE, C. F. GAUSS, voire même Daniel BERNOULLI. K. PEARSON a ensuite reconnu en 1920 que ce nom de *normal* était inadéquat - le terme *normal* avait pris un siècle plus tôt (1826) le sens de *habituel*, *typique* ou venant de *norme*, ayant pour signification *qui sert de règle, de modèle*, comme dans *école normale* de 1793 - et que ce nom « *a le désavantage de conduire les gens à croire que toutes les autres distributions de fréquences sont en un sens ou un autre* anormales. »[1]

I - Géométrie et courbes en cloche

Dès le début de la géométrie dans la Grèce antique, les géomètres se sont intéressés aux courbes, d'abord les plus simples : droite, polygones, cercle et plus généralement coniques, et ensuite aux courbes engendrées par des points mobiles. Parmi celles-ci, il y a des courbes en forme de cloche.

La plus ancienne courbe connue de ce genre est la *conchoïde de droite* ou *conchoïde de Nicomède* (NICOMÈDE est un géomètre de l'antiquité grecque du II^e^ siècle avant J.C.).

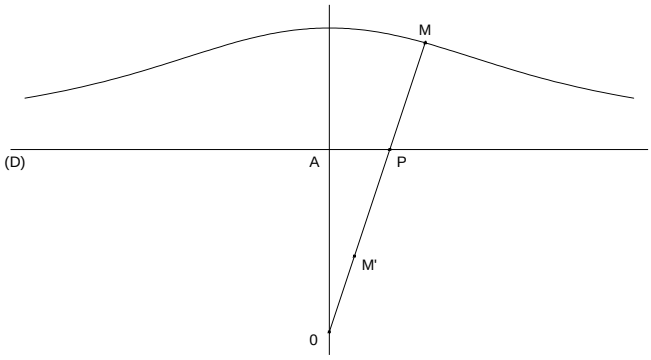
Une conchoïde est obtenue en prenant une droite (D), un point O non sur cette droite. A tout point P de (D), on associe les points M et M' situés de part et d'autre de P sur la droite (OP) à une distance b constante de P. Le lieu des points M et M' est une conchoïde de la droite (D).

C'est une courbe unicursale qui peut prendre diverses formes selon le rapport des longueurs b et OA, A étant le projecté orthogonal de O sur la droite (D). Cette courbe peut admettre une boucle et un point double ou un point de rebroussement en O.

Etudiée en géométrie cartésienne dans un système d'axes orthogonaux (Ox , Oy), c'est une courbe algébrique d'ordre 4 ; elle a une équation particulièrement simple en coordonnées polaires :

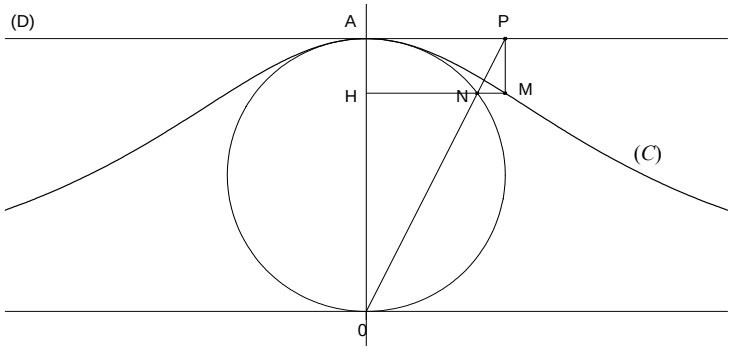
$$\rho = \frac{a}{\cos\theta} + b \text{ pour } M, \ \rho = \frac{a}{\cos\theta} - b \text{ pour } M', \text{ où } a = OA \text{ et } b = PM = PM'.$$

[1] Karl PEARSON : Note on the History of Correlation, *Biometrika*, **13**, 25-45 (1920) ; reproduit dans PEARSON et KENDALL, *Studies*..., vol. I.



Cependant cette courbe ne peut pas servir à définir une densité de probabilité car l'aire comprise entre la courbe et son asymptote, la droite (*D*), est infinie.

Une autre courbe de même forme a été étudiée moins anciennement ; elle est indiquée incidemment par Pierre DE FERMAT (1646) en application de son principe de *maximis & minimis*, construite par l'italien G. GRANDI en 1703, qui l'a nommée *versiera* en 1718, puis son étude a été faite un siècle après FERMAT par la mathématicienne italienne AGNESI[2] en 1748, d'où son nom *cubique versiera* ou *cubique d'Agnesi*.



Soit (*OA*) le diamètre d'un cercle, (*D*) la tangente au cercle en *A* ; la droite (*OP*) coupe le cercle en *N* et la tangente (*D*) en *P* ; sur le segment [*NP*], on construit un triangle rectangle *NPM* ayant son angle droit en *M* et dont les côtés [*PM*] et [*NM*] sont respectivement parallèles au diamètre (*OA*) et à la tangente (*D*). Lorsque le point *P* décrit la droite (*D*), le point *M* décrit la cubique (*C*).

[2] AGNESI, M.G. (1748) : *Instituzioni Analitiche...* (Milan). Voir l'article : Maria Gaetana Agnesi, *L'Ouvert* n° 96, sept. 1999, IREM de Strasbourg.

Une équation de la courbe (*C*) dans un système d'axes rectangulaires (*Ox* , *Oy*), en notant *a* la longueur du diamètre *OA*, est : $x^2y = a^2(a - y)$, ou encore $y = \frac{a^3}{x^2 + a^2}$.

L'aire entre la courbe et son asymptote, l'axe *Ox*, vaut quatre fois l'aire du cercle de diamètre *OA* ; cette courbe (à un coefficient convenable près) peut donc représenter une densité de probabilité. C'est la loi dite de Cauchy, $C(a)$, de paramètre $a > 0$, de densité $f_a(x) = \frac{1}{\pi} \frac{a}{x^2 + a^2}$.

Cette loi a été proposée par S.-D. POISSON, dans un mémoire de 1824[3], comme une distribution pour laquelle le Théorème-Limite Central ne s'applique pas. Elle fut étudiée par A. L. CAUCHY quelque vingt ans plus tard.

Passons maintenant à l'étude de la courbe en cloche dans un cadre statistique et probabiliste.

II - L'approche de J. BERNOULLI et ses améliorations : DE MOIVRE, BAYES, LAPLACE

Comme le notait déjà Jacques BERNOULLI dans son *Ars conjectandi* :

« pour former selon les règles des conjectures sur n'importe quelle chose il est seulement requis d'une part que les nombres de cas soient soigneusement déterminés, et d'autre part que soit défini combien les uns peuvent arriver plus facilement que les autres. Mais c'est ici enfin que surgit une difficulté, nous semble-t-il : cela peut se voir à peine dans quelques très rares cas et ne se produit presque pas en dehors des jeux de hasard que leurs premiers inventeurs ont pris soin d'organiser en vue de se ménager l'équité, de telle sorte que fussent assurés et connus les nombres de cas qui doivent entraîner le gain ou la perte, et de telle sorte que tous ces cas puissent arriver avec une égale facilité. En effet lorsqu'il s'agit de tous les autres résultats, dépendant pour la plupart soit de l'oeuvre de nature soit de l'arbitre des hommes, cela n'a pas du tout lieu. ... »,

lorsqu'on désire faire une application du « *calcul des chances* » aux « *affaires civiles, morales et économiques* », il faut pouvoir attribuer une valeur à la probabilité de chaque cas possible. Que faire si l'on n'est pas dans une situation où on peut percevoir quelque part une uniformité, i.e. des cas également possibles par un principe de raison insuffisante ? En suivant encore BERNOULLI :

« Mais à la vérité ici s'offre à nous un autre chemin pour obtenir ce que nous cherchons. Ce qu'il n'est pas donné d'obtenir a priori *l'est du moins* a posteriori*, c'est-à-dire qu'il sera possible de l'extraire en observant l'issue de nombreux exemples semblables ; car on doit présumer que, par la suite, chaque fait peut arriver et ne pas arriver dans le même nombre de cas qu'il avait été constaté*

[3] POISSON, S. D. (1824) : Sur la probabilité des résultats moyens des observations... *Mémoires de l'Académie des Sciences de Paris*.

auparavant, dans un état de choses semblables, qu'il arrivait ou n'arrivait pas... Cette manière empirique de déterminer par expérience les nombres de cas n'est ni neuve ni insolite. ... tout être des plus stupides, par je ne sais quel instinct naturel, par lui-même et sans le guide d'aucun enseignement (chose absolument admirable) tient pour évident que, plus on aura recueilli de nombreuses observations de ce genre, moins grand sera le danger de s'écarter du but. Or, bien que cela soit naturellement connu de tous, la démonstration qui permet de le tirer des principes de l'art n'est pas du tout répandue ... »

et encore,

« je voudrais que le rapport entre les nombres de cas, que nous entreprenons de déterminer expérimentalement, ne fût pas pris de façon nette et sans partage ..., mais je voudrais que le rapport fût admis dans une certaine latitude, c'est-à-dire compris entre une paire de limites, pouvant être prises aussi approchées qu'on voudra. »[4],

en langage moderne, BERNOULLI dit qu'il est bien connu que la fréquence observée est une estimation ponctuelle de la proportion inconnue et que le *degré de confiance* d'un intervalle augmente avec le nombre d'observations, et c'est ce qu'il se propose de démontrer.

L'étude du théorème de J. BERNOULLI a été très bien faite ailleurs[5]; je ferai juste quelques remarques. Il a été dit que J. BERNOULLI n'avait pas publié son ouvrage *Ars conjectandi* parce que son résultat est en pratique inapplicable en raison du grand nombre d'observations nécessaires pour avoir une probabilité proche de la certitude que la fréquence observée ne s'écarte pas trop de la proportion initiale. Un autre motif est peut-être que tout le calcul est basé sur cette proportion inconnue et ne peut donc servir pour évaluer la *probabilité* que l'écart entre la fréquence observée et la proportion inconnue soit inférieur à une quantité donnée.

D'après ces quelques citations, on voit que J. BERNOULLI accordait beaucoup d'importance au passage d'une fréquence observée à une proportion inconnue et à l'évaluation du *degré de confiance* d'un intervalle contenant la proportion inconnue, c'est-à-dire le problème de la probabilité inverse ; le théorème de Bernoulli se rapporte au problème direct : passage de la proportion (supposée donnée) à des probabilités concernant la fréquence qu'on observerait sur des expérimentations, c'est-à-dire, en langage moderne, la convergence en probabilité[6] de la suite des

---

[4] BERNOULLI, Jakob (1713) : *Ars conjectandi* ; trad. de la partie 4 dans MEUSNIER, N. (1987) : *Jacques Bernoulli et l'Ars conjectandi*, IREM de Rouen, extraits des pp. 40-46 (pp. 223-226 de *A.C.*).

[5] Voir N. MEUSNIER : *Jacques Bernoulli et l'Ars conjectandi*, *op. cit.*, et N. MEUSNIER : Argumentation et démonstration : A quoi sert la démonstration de la "Loi des grands nombres" de Jacques Bernoulli (1654-1705), dans *La démonstration mathématique dans l'histoire*, 7ème colloque Epistémologie et Histoire des mathématiques, IREM de Besançon, 1989.

[6] Le terme *convergence en probabilité* a été introduit par CANTELLI en 1916.

fréquences vers la proportion dans l'urne. Cette difficulté que BERNOULLI a dû percevoir est analogue à celle de l'établissement d'un intervalle de confiance.

En dépit du caractère insatisfaisant de son théorème pour J. BERNOULLI, celui-ci a mis en route une direction de recherche qui se développe encore de nos jours.

A. DE MOIVRE, dans son traité *The Doctrine of Chances* sur la théorie probabiliste[7], est arrivé à ce même problème, mais avec des outils mathématiques plus élaborés ; il a suivi une démarche analogue : l'étude des coefficients du développement du binôme. En 1733[8], il a obtenu une formule d'approximation de $\mathrm{Log}(n!)$, que son ami James STIRLING a précisée avec le coefficient $\frac{1}{2}\mathrm{Log}(2\pi)$[9]. DE MOIVRE montre que dans le développement de $(a+b)^n$, n grand, le Logarithme du rapport d'un terme au plus grand terme, situé à une distance inférieure ou égale à l de ce plus grand terme, est approximativement égal à $-\frac{(a+b)^2}{2abn} \times l^2$. Il fait alors la somme de tous les termes situés à une distance inférieure ou égale à l du plus grand terme, et calcule ensuite la valeur de la probabilité lorsque $l = s\sqrt{n}$ pour quelques valeurs de s, par un développement en série de l'exponentielle. Il indique d'ailleurs que $\sqrt{n}$ est le « Modulus *par lequel nous devons régler notre estimation* » (p. 248).

Cependant DE MOIVRE paraît considérer son résultat uniquement comme une approximation numérique des probabilités à partir de séries, la première mention de loi continue étant due à SIMPSON (1757). Avec son approximation des factorielles, DE MOIVRE arrivait effectivement à une amélioration notable par rapport à ce qu'avait obtenu J. BERNOULLI : le nombre d'observations nécessaires pour que la probabilité que le nombre de succès tombe dans un intervalle fixé soit proche de 1 était ramené à un nombre *raisonnable*.

Dans le corollaire précédant cet appendice, DE MOIVRE écrit (p. 242) :

« De cela il s'ensuit que si, après avoir pris un grand nombre d'expériences, il a été perçu que les survenances et les échecs ont été de très près dans une certaine proportion ... il peut être conclu avec sûreté que les probabilités d'une survenance ou d'un échec à un moment quelconque assigné seront de très près en cette proportion, et plus le nombre d'expériences a été grand, d'autant plus près de la vérité seront les conjectures qui en sont dérivées. »,

---

[7] DE MOIVRE, Abraham : *The Doctrine of Chances* (1ère éd., 1718 ; 2ème éd., 1738 ; 3ème éd., 1756). Les références de pages sont relatives à la 3ème éd. reproduit (1967 ; 2000), New-York, Chelsea.

[8] DE MOIVRE, Abraham (1733) : Note *Approximatio ad Summam Terminatorum Binomii* $\overline{\mathrm{a}+\mathrm{b}}^n$ *in Seriem expansi*, traduite en anglais par De Moivre : *A Method of approximating the Sum of the Terms of the binomial* $\overline{\mathrm{a}+\mathrm{b}}^n$ *expanded into a Series*, dans la 2éme éd. (1738) et la 3ème éd. (1756) de *The Doctrine of Chances*.

[9] Cette formule devrait être appelée en toute justice *formule de De Moivre-Stirling* :

$$\ln(n!) \sim n\ln\left(\frac{n}{\mathrm{e}}\right) + \frac{1}{2}\ln(n) + \frac{1}{2}\ln(2\pi).$$

DE MOIVRE note aussi, dans une remarque qui suit cet appendice (p. 251), que la convergence permet de déterminer expérimentalement une proportion inconnue, c'est-à-dire une solution asymptotique du problème de la probabilité inverse :

« Remark II

As, upon the Supposition of a certain determinate Law according to which any Event is to happen, we demonstrate that the Ratio of Happenings will continually approach to that Law, as the Experiments or Observations are multiplied: so, conversely*, if from numberless Observations we find the Ratio of the Events to converge to a determinate quantity, as to the Ratio of P to Q; then we conclude that this Ratio expresses the determinate Law according to which the Event is to happen. »*[10]

DE MOIVRE, comme J. BERNOULLI, notait que la question importante était celle de la probabilité inverse, mais la difficulté déjà rencontrée par J. BERNOULLI n'était pas surmontée : *pour un nombre fini d'observations données, que dire sur la valeur de la proportion inconnue* ? Une autre approche était nécessaire, faite ensuite par T. BAYES (1763) et P. S. LAPLACE (1774).

Quoique n'étant pas directement reliées avec la loi normale, je mentionnerais ici les premières recherches faites sur le problème de la probabilité inverse (l'estimation *a posteriori*) initié par J. BERNOULLI.

Le premier mémoire publié est l'*Essay* posthume de Thomas. BAYES en 1763[11]. Il se proposait d'évaluer, au moyen de son dispositif ingénieux de boules lancées sur une table carrée et dont la position d'arrêt est au hasard sur la table, la probabilité que la première boule lancée se soit arrêtée dans une bande donnée (c'est-à-dire une probabilité inférieure et une probabilité supérieure), connaissant la position d'arrêt de boules lancées après. Ce dispositif permettait d'éviter de parler de la *probabilité* qu'une probabilité soit dans un intervalle. Pour une étude détaillée, voir la traduction annotée de l'*Essay* de BAYES par J. P. CLÉRO.

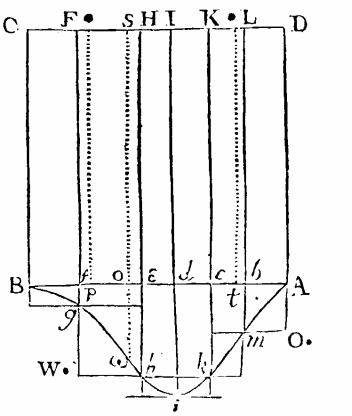
Sur le graphique suivant, sous le dessin de la table carrée ABCD apparaît une courbe en forme de cloche. La boule initiale ayant une position d'arrêt d'abscisse x, la courbe d'équation $y = Kx^p(1-x)^q$, où K est un coefficient adéquat, représente la

---

[10] « Comme, sous l'hypothèse qu'un événement doit se produire suivant une certaine loi, nous démontrons que le rapport de ses réalisations se rapprochera continuellement de cette loi quand les expériences ou les observations sont multipliées : inversement, si à partir d'innombrables observations nous obtenons que le rapport des événements converge vers une quantité déterminée, comme le rapport de P à Q, alors nous concluons que ce rapport exprime la loi selon laquelle cet événement doit se produire ». (le "rapport" considéré est celui du nombre des réalisations de l'événement à celui des apparitions de l'événement contraire, traduction de M. HENRY).

[11] BAYES, Thomas (1763) : An Essay towards solving a Problem in the Doctrine of Chances. *Philosophical Transactions*, London ; reproduit dans PEARSON and KENDALL, *Studies I* ; traduit par J. P. CLÉRO : *Cahiers d'Histoire et de philosophie des Sciences*, n° 18, 1988.

distribution des abscisses des positions d'arrêt des boules lancées après, p boules ayant une abscisse plus petite que x et q boules ayant une abscisse plus grande.



C'est une fonction dont EULER[12] a étudié l'intégrale $B(p\,,q) = \int_0^1 x^{p-1}(1-x)^{q-1}\,dx$ pour p et q strictement positifs, appelée fonction Bêta ou intégrale Eulérienne[13] de première espèce.

L'intégrale Eulérienne de seconde espèce ou fonction Gamma intervient aussi en calcul des probabilités, $\Gamma(x) = \int_0^1 t^{x-1}\,e^{-t}\,dt$. La fonction Γ généralise la factorielle : si n est un entier, $n! = \Gamma(n+1)$.

On trouve aussi cette fonction et intégrale Eulérienne dans le premier grand mémoire de P.-S. LAPLACE sur les probabilités[14]. Ce mémoire de 1774 de LAPLACE a, semble-t-il, été fait indépendamment de l'*Essay* de BAYES, qui a été mentionné en France seulement en 1780 par D'ALEMBERT et CONDORCET. LAPLACE, utilisant le concept de probabilité conditionnelle, énonce un principe[15] qui, écrit sous forme d'une formule, est appelé maintenant *formule de Bayes*. Le reprenant dans son *Essai philosophique* en sixième principe, il l'énonce en deux temps :

---

[12] EULER, Leonhard (1748-1773) : *Leonhardi Euleri Opera Omni*, série I, Opera mathematica, vols. 17 à 21, Teubner, Leipzig et Berlin, 1912-1932.

[13] Nom donné par LEGENDRE, ainsi que la notation Γ dans *Traité des fonctions elliptiques et des intégrales eulériennes*, 3 vols, 1825-1828.

[14] LAPLACE, P.S. (1774) : Mémoire sur la probabilité des causes par les événements, Œuvres complètes, VIII pp. 27-65.

[15] Dans le mémoire de 1774, ce principe est énoncé ainsi : "*Si un événement peut être produit par un nombre n de causes différentes, les probabilités de l'existence de ces causes prises de l'événement sont entre elles comme les probabilités de l'événement prises de ces causes, et la probabilité de l'existence de chacune d'elles est égale à la probabilité de l'événement prise de cette cause, divisée par la somme de toutes les probabilités de l'événement prises de chacune de ces causes.*"

« Chacune des causes, auxquelles un événement *observé peut être attribué, est indiquée avec d'autant plus de vraisemblance, qu'il est plus probable que cette cause étant supposée exister, l'événement aura lieu ; la probabilité de l'existence d'une quelconque de ces causes est donc une fraction dont le numérateur est la probabilité de l'événement, résultante de cette cause, et dont le dénominateur est la somme des probabilités semblables relatives à toutes les causes… »*

Ce qui peut se traduire, en notations contemporaines par la formule :

$$P_E(C_i) = \frac{P_{C_i}(E)}{\sum_i P_{C_i}(E)}$$

où l'on voit que LAPLACE, en vertu du principe de raison, en l'absence d'information sur les « possibilités » des causes, se place d'abord dans le cadre de leur équiprobabilité. Il ajoute cependant :

« *…Si ces diverses causes considérées* a priori *sont inégalement probables, il faut au lieu de la probabilité de l'événement, résultante de chaque cause, employer le produit de cette probabilité par la possibilité de la cause elle-même. C'est le principe fondamental de cette branche de l'Analyse des hasards, qui consiste à remonter des événements aux causes.* »

ce qui donne la formule dite « de Bayes » :

$$P_E(C_i) = \frac{P(C_i) \times P_{C_i}(E)}{\sum_i P(C_i) \times P_{C_i}(E)}$$

Dans son mémoire de 1774, il propose ce problème :

« PROBLÈME I. – Si une urne renferme une infinité de billets blancs et noirs dans un rapport inconnu, et que l'on en tire p + q billets dont p soient blancs et q soient noirs; on demande la probabilité qu'en tirant un nouveau billet de cette urne il sera blanc. »

Il dit que si on prend x pour représenter ce rapport inconnu, x étant un des nombres depuis 0 jusqu'à 1, la probabilité que x est le vrai rapport du nombre des billets blancs au nombre total des billets est par le principe précédent égale à :

$$\frac{x^p\,(1-x)^q\,dx}{\int x^p\,(1-x)^q\,dx},$$

l'intégrale étant prise de 0 à 1.

Il obtient alors que la probabilité qu'un nouveau billet tiré de l'urne sera blanc est égale à $\frac{p+1}{p+q+2}$. C'est l'estimation *a posteriori* Laplacienne de la probabilité d'un événement. La suite de ce Problème I porte sur la convergence quand p et q deviennent infinis ; j'y reviendrai dans le paragraphe consacré au théorème-limite central.

III - Les premiers pas : la théorie des erreurs

On s'était aperçu depuis longtemps que plusieurs mesurages d'une caractéristique d'un même objet céleste donnaient souvent des valeurs différentes entre elles et les astronomes préconisaient (e.g. PTOLÉMÉE, IIe siècle, et KÉPLER, fin du XVIe siècle) de prendre comme mesure de la caractéristique la moyenne arithmétique ou la moyenne élaguée (en supprimant des valeurs extrêmes) des observations. Avec l'augmentation considérable de la précision des instruments de visée (quadrant ou quart de cercle couplé avec deux lunettes astronomiques, des micromètres et des mires réticulaires) à la fin du XVIIe siècle et au début du XVIIIe siècle, la variabilité des mesurages d'un même objet céleste (ou terrestre en géodésie) devenait plus flagrante. La question de *comment attribuer au mieux une valeur pour la caractéristique d'un objet quand les observations donnent des résultats différents* se posait à ceux qui pratiquaient des mesures précises (astronomes, physiciens, etc.).

COTES (1722)[16] indique que la moyenne de plusieurs observations est la valeur la plus probable, mais il ne dit pas comment il est arrivé à cette règle. Lors des grandes expéditions françaises (1735-1744) concernant la mesure de l'arc de méridien, la mise en évidence des aberrations et turbulences de l'atmosphère avaient conduit à utiliser des méthodes probabilistes, mais celles-ci n'ont pas été publiées[17].

Quelques années plus tard, à la suite de la mesure d'un arc de méridien en Italie, MAIRE et BOSCOVICH[18] (1755) constatent que les paires prises parmi les cinq longueurs d'arcs de méridien mesurés à des latitudes différentes donnent des valeurs différentes de l'ellipticité de la terre. BOSCOVICH en 1757 indique une méthode pour combiner ces observations. Le critère qu'il propose pour déterminer les coefficients d'une relation linéaire entre deux quantités (le sinus de la latitude et la longueur d'arc de méridien de 1°) est basé sur deux conditions (utilisées implicitement depuis longtemps) :

- Les sommes des corrections positives et des corrections négatives devront être égales.
- La somme de toutes les corrections, positives et négatives, devra être aussi petite que possible.

---

[16] COTES, R. (1722) : Aestimatio errorum in mixta mathesi,... in *Harmonia mensurarum*.

[17] Voir l'article de B. BRU : *Laplace et la critique probabiliste des mesures géodésiques*, dans H. LACOMBE et P. COSTABEL (1988).

[18] MAIRE, Christopher et BOSCOVICH, Roger (1755) : *De Litteraria Expeditione per Pontificiam ditionem ad dimetiendas duas Meridiani gradus* .... Romae. Traduit dans MAIRE et BOSCOVICH, (1770) : *Voyage Astronomique et Géographique dans l'État de l'Église, entrepris par l'Ordre et sous les Auspices du Pape Benoît XIV, pour mesurer deux degrés du méridien, et corriger la Carte de l'État ecclésiastique*, Paris. Voir aussi l'article de EISENHART dans Kendall and Plackett, éds, *Studies II*.

BOSCOVICH reviendra plus tard (en 1770) sur cette méthode et indiquera comment la mettre en œuvre de façon géométrique.

Les deux questions soulevées : la meilleure valeur à attribuer à une mesure à partir de diverses observations d'un même objet (le problème de l'estimation) et la combinaison d'observations ayant une relation fixée entre elles – ici, dans la question de la figure de la Terre, la détermination de l'ellipticité à partir des diverses longueurs d'arcs de méridien – (le problème de la régression) sont fondamentales en statistique et les recherches sur ces sujets continuent encore (à un niveau évidemment plus général).

Les scientifiques français engagés dans la mesure d'un arc de méridien n'ayant pas indiqué dans leurs comptes rendus d'expédition les méthodes statistiques utilisées pour la correction des erreurs[19], c'est au mathématicien anglais, T. SIMPSON (1756-57), que l'on doit la première publication sur ce sujet[20].

SIMPSON remarque tout d'abord que ce qui importe ce ne sont pas les observations par elles-mêmes, mais les écarts entre ces observations (supposées indépendantes) et la vraie valeur (inconnue) de la position de l'objet céleste étudié, i.e. l'erreur commise sur la mesure, d'où le nom de *théorie des erreurs*. Il effectue ensuite une analyse probabiliste de ces erreurs selon le même schéma que celui de Jakob BERNOULLI : les erreurs peuvent être positives ou négatives, les erreurs positives et négatives de même grandeur ayant la même probabilité de survenir, avec une distribution discrète rectangulaire puis triangulaire centrée en 0.

Le fait de prendre l'erreur plutôt que la valeur de l'observation permettait de dépasser la limitation imposée dans le traitement de J. BERNOULLI, puis de DE MOIVRE, de l'estimation d'une proportion où la distribution de la différence entre fréquence empirique et la proportion inconnue dépend de la véritable valeur de cette proportion inconnue.

Le second pas important fait par SIMPSON est de considérer, après des distributions discrètes, des distributions continues (rectangulaires puis triangulaires) car les erreurs sont susceptibles de prendre des valeurs qui ne sont pas seulement des valeurs entières, mais toutes les valeurs sur un domaine borné.

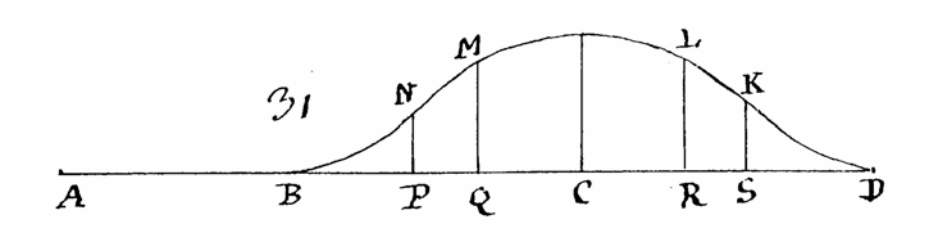
Les résultats obtenus par SIMPSON, cependant, sont modestes : utilisant les fonctions génératrices développées par DE MOIVRE, il montre que l'erreur sur la moyenne de plusieurs observations est plus petite que sur chaque observation prise isolément, ce qui allait néanmoins à l'encontre de l'opinion de certains expérimentalistes qui affirmaient qu'il valait mieux une seule observation soigneusement réalisée.

---

[19] MAUPERTUIS (1738) : *La Figure de la Terre*. BOUGUER (1749). La CONDAMINE (1751).

[20] SIMPSON, Thomas (1757) : An Attempt to show the advantage of taking the mean of a number of observations in practical astronomy, *Miscellaneous Tracts...*, London.

Mais comme dans le cas du théorème de Jakob BERNOULLI qui est, d'après ses propres mots, une démonstration mathématique d'un fait bien connu intuitivement par expérience, cet article de SIMPSON est le point de départ des recherches sur la théorie des erreurs. Diverses formes de loi de facilité des erreurs sont proposées, e.g. LAMBERT en 1760 propose une courbe en cloche (à domaine borné) obtenue probablement à partir d'un histogramme lissé correspondant à de nombreuses expériences[21].



Des mathématiciens de plus grand renom et de plus grande envergure s'intéressent alors au problème de la théorie des erreurs. En reprenant les cas de SIMPSON, D. BERNOULLI en 1777[22], et LAGRANGE en 1773[23] soulignent que « *il n'y a aucun doute que les petites erreurs ont lieu plus souvent que les grandes* », mais ils considèrent encore que les erreurs sont bornées, par exemple D. BERNOULLI propose une distribution des erreurs en arc de cercle.

LAGRANGE a utilisé aussi une extension des fonctions génératrices de DE MOIVRE (ce qui sera nommé la transformation de Laplace) pour déterminer la distribution de la moyenne, mais il lui manquait une formule d'inversion pour aboutir[24]. C'était aussi le cas pour les recherches de LAPLACE (1774-1786) qui ont débouché seulement à partir de 1809-1810 avec sa formule d'inversion pour la transformation de Fourier (voir ci-après).

---

[21] Graphique extrait de LAMBERT, Jean Henri (1760), *photometria sive de mensura…* ; d'après BOYÉ, A. et LEFORT, X. (1996) : De Cassini à Gauss : du calcul d'erreurs aux probabilités dans *Actes de la 6ème Université d'été sur l'histoire des mathématiques*, 1995, IREM de Besançon.

[22] BERNOULLI, Daniel (1778). Dijudicatio maxime probabilis plurium observationum discrepantium... *Novi Comm. Ac. Sc. Imp. Petrop* (pour 1777). Traduit *The most probable choice...* dans PEARSON et KENDALL, *Studies I*. Ici, D. BERNOULLI introduit aussi la méthode du maximum de vraisemblance.

[23] LAGRANGE, J. L. (1773) : Mémoire sur l'utilité de la méthode de prendre le milieu entre les résultats de plusieurs observation. *Misc. Taurinensia* (1770/73), publié en 1776. *Oeuvres*, tome 2, Paris, 1868.

[24] Cette question est à rapprocher du problème des cordes vibrantes étudié par D'ALEMBERT, Daniel BERNOULLI et Leonhard EULER dans les années 1745-55. Daniel BERNOULLI avait bien vu que des fonctions trigonométriques intervenaient dans la résolution de l'équation des cordes vibrantes, mais il lui manquait le passage de la fonction aux coefficients.

LAPLACE dans son mémoire (1774) codifie les conditions que doit vérifier une loi de facilité des erreurs ; citons son

« PROBLÈME III. – Déterminer le milieu que l'on doit prendre entre trois observations données d'un même phénomène :

... la loi suivant laquelle cette vraisemblance diminue à mesure que l'observation s'éloigne de la vérité nous est inconnue. Supposons donc (fig. 2) que le point V soit le véritable instant du phénomène ; ... et en nommant x l'abscisse VP, et y l'ordonnée correspondante PM, nous représenterons l'équation par celle-ci : $y = \varphi(x)$.

Fig. 1.

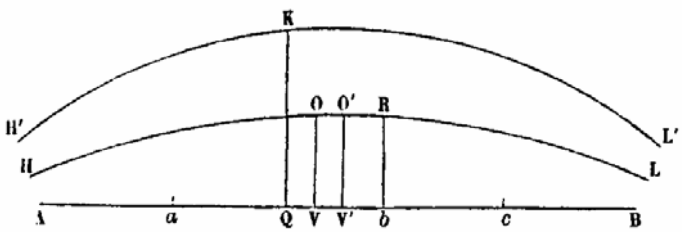
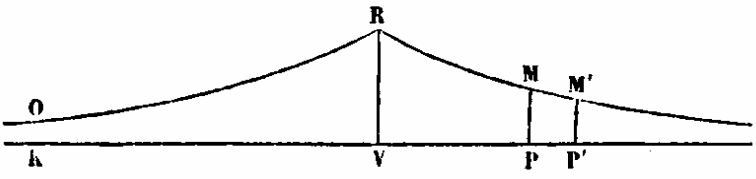


Fig. 2.



Or voici les propriétés de cette courbe :

1° elle doit être partagée en deux parties entièrement semblables par la droite VR, car il est tout aussi probable que l'observation s'écartera de la vérité à droite comme à gauche ;

2° elle doit avoir pour asymptote la ligne KP, parce que la probabilité que l'observation s'éloigne de la vérité à une distance infinie est évidemment nulle ;

3° l'aire entière de cette courbe doit être égale à l'unité, puisqu'il est certain que l'observation tombera sur un des points de la droite KP. » [25]

puis il donne quelques graphes de répartition possibles (lois de facilité) qui ont un domaine infini.

En utilisant le critère des moindres écarts de BOSCOVICH, c'est-à-dire, « *l'instant tel qu'en le prenant pour milieu, la somme des erreurs à craindre, multipliées par*

[25] LAPLACE, P. S. (1774) : *Œuvres complètes*, Les citations sont des pp. 42-46.

leur probabilité, soit un minimum » et avec l'hypothèse supplémentaire « *Or, comme nous n'avons aucune raison de supposer une autre loi aux ordonnées qu'à leurs différences ...*», LAPLACE montre que cette loi des erreurs doit être la loi dite *première loi de Laplace*, appelée aussi loi exponentielle bilatère, dont il a donné un graphe dans sa *fig.* 2.

On peut remarquer que cette loi de facilité proposée par LAPLACE est la première à avoir un domaine non borné, c'est-à-dire que les erreurs peuvent éventuellement avoir des valeurs très grandes, voire infinies. Mais quand LAPLACE reprend ce problème de la théorie des erreurs en 1781[26], il propose une loi de facilité à domaine borné, $\frac{1}{2a}\log\left(\frac{a}{|x|}\right)$ pour $|x| < a$. On doit noter aussi que ces différentes propositions de loi de facilité des erreurs n'ont pas de justification théorique, hormis peut-être – via son hypothèse supplémentaire – la première loi de Laplace.

IV - Le théorème de De Moivre-Laplace

La première forme, et la plus simple, de ce qui est appelé maintenant le théorème-limite central concerne la convergence (en loi) de la loi binomiale vers la *loi des erreurs* (nom proposé par WILSON[27] pour obvier au défaut du nom *loi normale* signalé par K. PEARSON en 1920).

Nous avons vu que A. DE MOIVRE avait obtenu une approximation de la probabilité que le nombre de succès sur n épreuves indépendantes, chacune ayant la même probabilité p de succès, ne s'écarte pas de np de plus de $\alpha\sqrt{n}$, par une somme de termes de la forme $\exp\left(-\frac{(a+b)}{2abn}\times l^2\right)$. S'il avait remarqué un résultat de MACLAURIN[28] de 1742 (mais DE MOIVRE était alors âgé de 75 ans) sur le passage d'une somme à une intégrale, formule connue sous le nom de *formule d'Euler*[29], il serait arrivé à $k\int e^{-t^2}\,dt$, où k est un coefficient adéquat et l'intégrale est à prendre entre deux bornes à calculer.

D. BERNOULLI (1770/71)[30], redémontre le théorème de De Moivre, l'approximation de probabilités liées à la loi binomiale pour des grands échantillons, et montre qu'intervient la fonction e^{-kx^2} dont il donne une petite table pour quelques valeurs de x.

---

[26] LAPLACE, P. S. (1781) : Mémoire sur les probabilités, *Œuvres complètes*, IX, pp. 383-485.

[27] WILSON, E.-B. (1923) : First and Second Laws of Error. *Quaterly publications of the Amer. Stat. Assoc.*

[28] MACLAURIN, Colin (1742) : *Treatise of Fluxions*.

[29] EULER, Leonhard. (1770) : *Novi Comm. Acad. Sc. Imp. Petrop.,* Vol. XIV.

[30] BERNOULLI, Daniel (1770-71). Mensura sortis ad fortuitam... *Novi Comm. Acad. Sc. Imp. Petrop*.

Revenons maintenant au mémoire de 1774 de LAPLACE et la suite de son Problème I. Il prend les nombres p de billets blancs et q de billets noirs infinis et, comme J. BERNOULLI et DE MOIVRE, il cherche la probabilité que la proportion inconnue de billets blancs dans l'urne soit entre $\frac{p}{p+q} - w$ et $\frac{p}{p+q} + w$. Il prend le changement de variable $x = \frac{p}{p+q} + z$ dans l'intégrale eulérienne incomplète du numérateur et fait une approximation des fonctions à intégrer, e.g. $\left(1 + \frac{p}{p+q}\right)^p$. Il aboutit à une intégrale de la forme $K \int e^{-K'z^2} dz$. En citant EULER (« *voir les* Institutions du Calcul différentiel *de M. Euler*[31] »), il fait l'approximation de la factorielle[32] et le calcul de l'intégrale, qui équivaut au résultat[33] $2 \int_0^\infty e^{-y^2} dy = \sqrt{\pi}$.

Ce résultat est ainsi beaucoup plus utilisable que la forme de DE MOIVRE, et est maintenant appelé *théorème de De Moivre-Laplace*[34]. Cependant, cette fonction n'est pas explicitement considérée ici comme une densité de probabilité : un graphe de même forme apparaît bien dans sa figure 1 du Problème III (voir ci-dessus) mais la seule *loi de facilité* étudiée par LAPLACE est la Première loi de Laplace (la loi exponentielle bilatère). Il a été proposé comme explication que LAPLACE, cherchant à trouver le meilleur milieu à prendre entre trois observations, est arrivé avec la loi exponentielle bilatère à une équation du 15ème degré, et qu'il a dû penser que la loi de forme plus compliquée en e^{-kx^2} conduirait à des calculs encore plus inextricables. LAPLACE a repris ce théorème dans son mémoire de 1781 mais avec une autre démonstration et la validité des approximations est mieux justifiée.

Dans un mémoire de 1785[35], LAPLACE se propose de réduire en séries convergentes des intégrales composées d'un grand nombre de termes et de facteurs

---

[31] EULER, Leonhard (1755) : *Institutiones calculi differentialis*, Berlin. (1769) : *Institutiones calculi integralis*.

[32] C'est la formule de De Moivre-Stirling. EULER a publié ce résultat dans *Comm. Ac. Sc. Petrop.,* (pour 1730-31).

[33] Certains auteurs appellent cette intégrale, *intégrale de Gauss*, mais cette appellation est tout à fait inappropriée, car en 1774, C. F. GAUSS n'était pas né ! D'ailleurs GAUSS lui-même mentionne dans *Méthode des moindres carrés* le résultat $\int_0^\infty e^{-t^2} dt = \frac{\sqrt{\pi}}{2}$ comme le « *beau théorème de M. La Place* », cette attribution est elle-même inadéquate, comme LEGENDRE l'a fait remarquer, et LAPLACE dans son mémoire de 1774 attribue ce résultat à Leonhard EULER : « *Voir le* Calcul intégral *de M. Euler*. »

[34] Il pourrait à juste titre être appelé *Théorème de De Moivre-D. Bernoulli-Laplace*.

[35] LAPLACE, P. S. (1785). Mémoire sur les approximations des formules qui sont fonctions de très grands nombres. *Mémoires de l'Académie royale des Sciences de Paris*, année 1782 ; 1785 ; *Œuvres complètes*, X, pp. 209-291.

et il indique que ces « *suites renferment des quantités transcendantes qui, le plus souvent, se réduisent à celle-ci :* $\int dt\, e^{-t^2}$ ». Il donne des formules d'approximation pour cette intégrale de 0 à T pour T petit, et de T à ∞ pour T grand. La suite du mémoire (1786)[36] est intitulée *Application de l'analyse précédente à la théorie des hasards* et commence par un préambule qui est repris dans l'*Essai* (pp. 32-35). Il écrit (p. 298) :

« Il est donc indispensable d'avoir un moyen simple d'obtenir la loi suivant laquelle la probabilité d'un résultat indiqué par les observations croît avec elles ».

Mais, malgré sa dextérité à manier les approximations, LAPLACE n'a pu résoudre la détermination de la loi d'une somme de quantités aléatoires un peu plus générales que celles données par un schéma de Bernoulli, parce que cela fait intervenir un grand nombre d'itérations d'intégrales et il se limite à des applications portant sur des répétitions de schémas de Bernoulli et obtient d'une manière différente les résultats déjà obtenus en 1774 et 1781. Il ne peut traiter les cas comme les erreurs de mesure qui peuvent prendre chacune une valeur quelconque dans un intervalle (c'est-à-dire une infinité de cas possibles). LAPLACE devait avoir cependant une idée, quoique confuse, de ce qui allait devenir le théorème-limite central car il écrit (p. 305) :

« L'intégrale $\int \mathrm{dt}\, \mathrm{e}^{-\mathrm{t}^2}$ *se rencontre fréquemment dans cette analyse et, par cette raison, il serait très utile de former une Table de ses valeurs, depuis* $t = \infty$ *jusqu'à* $t = 0$ *... ».*

Karl PEARSON voyait dans cette phrase l'origine de la loi normale[37]. Cette loi des erreurs était probablement utilisée en astronomie car KRAMP[38] a publié en 1799 (sur l'instigation de LAPLACE) une petite table donnant des valeurs de $\int_0^t e^{-\tau^2} d\tau$.

[36] LAPLACE, P. S. (1786). Suite du mémoire précédent, *Œuvres complètes,* X, pp. 295-338.

[37] PEARSON, Karl (1920) : Note on the History of Correlation, *Biometrika* ; reproduit dans Pearson et Kendall, *Studies I*.

[38] KRAMP, C., (1799) : *Analyse des réfractions astronomiques et terrestres*, Strasbourg et Leipzig.

V - La méthode des moindres carrés

LAPLACE va se consacrer ensuite pendant la fin du XVIII[e] siècle à son œuvre la plus importante, la *Mécanique Céleste*[39]. C'est d'ailleurs de l'astronomie que va venir le déclic qui permettra d'arriver à une solution dans la théorie des erreurs. A. M. LEGENDRE publie en 1805 son livre[40] qui expose une méthode pour résoudre des systèmes d'équations linéaires comportant plus d'équations que d'inconnues, systèmes appelés surdéterminés, auxquels on arrive dans le cas des observations indirectes, comme par exemple pour la détermination des paramètres du sphéroïde terrestre à partir des mesures de longueur d'un arc de méridien de 1° à différentes latitudes. Pour l'observation *i*, on a une

« équation de la forme $E_i = a_i + b_i x + c_i y + f_i z + etc.$ *où* a_i, b_i, c_i, f_i, *etc. sont des coefficients connus qui dépendent de la valeur de l'observation, x, y, z, etc. sont des inconnues qu'il faut déterminer par la condition que* E_i *se réduise, pour chaque équation, à une quantité nulle ou très petite. »*

Le principe que propose LEGENDRE est de « *rendre* minimum *la somme des quarrés des erreurs* » E_i. C'est donc un principe géométrique lié à la structure euclidienne de l'espace. Par différentiation de $\sum E_i^2$ par rapport à chaque inconnue, LEGENDRE obtient les équations normales puis donne des estimations des inconnues, mais le cadre probabiliste reste très flou. Un américain, R. ADRAIN, en étudiant cette méthode des moindres carrés, a atteint en 1808 la loi normale par une route différente de celle utilisée par GAUSS, mais son mémoire[41] est passé inaperçu.

GAUSS (1809)[42] donne la première discussion, dans un cadre probabiliste, d'un système d'équations de condition linéaires, du type considéré par LEGENDRE ; il montre que si la moyenne des erreurs d'observations indépendantes trouvée par la méthode des moindres carrés est la valeur la plus probable, alors la loi des erreurs (supposée continue) est la loi normale.

---

[39] LAPLACE, P. S. (1779-1825) : *Mécanique Céleste*, vol. I et II, 1799 ; Vol. III, 1803 ; vol. IV, 1805, vol. V, 1825. *Œuvres complètes,* II à IV.

[40] LEGENDRE, Adrien Marie (1805) : *Nouvelles méthodes pour la détermination des orbites des comètes*, Paris. (Appendice *Sur la méthode des moindres quarrés*.)

[41] ADRAIN, Robert (1808) : Research concerning the probabilities of the errors which happen in making observations... *The Analyst; or Mathematical Museum*.

[42] GAUSS, C. F. (1809) : *Theoria motus corporum cœlestium...* ; (1821-26) *Theoria combinationis...* Extraits dans *Méthode des moindres carrés. Mémoires sur la combinaison des observations*, Trad. J. BERTRAND, Paris, 1855, reproduit dans *Friedrich Gauss. Méthodes des moindres carrés*, IREM de Paris VII, 1996.

VI - Le théorème-limite central

J'utilise ici le nom *théorème-limite central* comme traduction de *zentraler Grenzwertsatz*, nom donné par G. POLYÁ, en 1920[43] et qui a été traduit en anglais par *central limit theorem* et ensuite traduit en français sous diverses formes ; ici l'adjectif *central* réfère à *théorème-limite* et non à *limite*.[44]

Indiquons ici le déclic qui a incité LAPLACE à revenir sur le problème de la théorie des erreurs. En 1807, Joseph FOURIER[45] a présenté à l'Académie des Sciences de Paris un mémoire sur la théorie de la chaleur (non publié, remanié et prix de l'académie en 1809, publié en 1822[46]) dans lequel il étudiait les propriétés de séries trigonométriques, et le développement d'une fonction périodique en série trigonométrique (dite maintenant en série de Fourier).

Ce n'était pas la première introduction de séries trigonométriques pour résoudre un problème venant de la physique. Dans les années 1745-1755, D'ALEMBERT, Daniel BERNOULLI et Leonhard EULER s'étaient attaqués au problème des cordes vibrantes et Daniel BERNOULLI avait eu l'idée de représenter les fonctions solutions de l'équation des cordes vibrantes sous forme d'une série trigonométrique, pour des raisons physiques.

Avec ces nouveaux outils, LAPLACE revient sur le problème qu'il avait essayé de résoudre en 1785, et cette fois ses efforts sont couronnés de succès (1810-1811[47]). Il part des fonctions génératrices de DE MOIVRE pour les variables à valeurs entières et de leurs propriétés pour les sommes de quantités aléatoires indépendantes. Si X est une variable aléatoire[48] à valeurs entières, de loi de probabilité $\{p_x\}$, $x \in \mathbb{N}$, sa fonction génératrice est $g_X(t) = \sum_{x\geq 0} p_x\, t^x$ et la probabilité $P(X = k)$ est le coefficient du terme en t^k dans le développement en série de la

[43] POLYA, G. Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem, *Math Zeitschrift*, t.**8**, (1920), pp. 171-181.

[44] Si POLYÁ avait utilisé l'adjectif *fondamental* au lieu de *central*, l'expression *théorème limite fondamental* aurait eu une signification adaptée, mais je pense qu'on n'aurait pas vu de locution comme *théorème de la limite fondamentale* !

[45] FOURIER était alors préfet de l'Isère.

[46] FOURIER, Joseph (1822) : *Théorie Analytique de la Chaleur* (réédition, J. Gabay, 1988).

[47] LAPLACE, P. S. : Mémoire sur les approximations des formules qui sont fonctions de très grands nombres, et sur leur application aux probabilités. 1810 ; *Œuvres complètes,* XII, pp. 301-349. Supplément au mémoire précédent. 1811 ; *Œuvres complètes,* XII, pp. 349-353 ; Mémoire sur les intégrales définies, et leur application aux probabilités, et spécialement à la recherche du milieu qu'il faut choisir entre les résultats des observations, 1811, *Œuvres complètes,* XII, pp. 357-412 ; voir aussi *T.A.P.*, Livre II, n° 20, 25.

[48] Ce terme *variable aléatoire* a été introduit par CHEBYSHEV dans la seconde moitié du XIX^{e} siècle.

fonction génératrice g_X. Si X_1, ..., X_n sont des variables aléatoires indépendantes de même loi que X, la fonction génératrice de la somme $S = \sum_{j=1}^{n} Xj$ est $g_S(t) = \left(g_X(t)\right)^n$.

Dans quelques cas simples, on peut trouver la probabilité $P(S = k)$ par la méthode ci-dessus, mais en général le problème inverse du passage de la fonction génératrice aux probabilités $P(S = k)$ est impraticable. Dans le cas d'une fonction f périodique de période T exprimée en une série trigonométrique

$$f(x) = \sum_{n \in \mathbb{N}} \left[a_n \cos\frac{2\pi nx}{T} + b_n \sin\frac{2\pi nx}{T} \right]$$

les coefficients a_n et b_n sont trouvés par la formule d'inversion de Fourier, par exemple : $a_n = \frac{2}{\mathrm{T}} \int_0^T f(x) \cos\frac{2\pi nx}{T}\, dx$.

LAPLACE va combiner les deux méthodes : dans la fonction génératrice, il remplace la puissance t^x par l'exponentielle complexe, c'est-à-dire la fonction $\varphi_X(w) = \sum_{x \geq 0} p_x \, \mathrm{e}^{xw\sqrt{(-1)}}$ (et se ramène au cas trigonométrique) et la formule d'inversion lui permet de trouver les probabilités limites pour la somme d'un grand nombre de variables indépendantes. Il obtient ainsi une méthode, qu'il nomme des fonctions génératrices, appelée maintenant des fonctions caractéristiques, qui sera reprise et développée un siècle plus tard, notamment par Paul LÉVY[49] (qui montrera que s'il y a convergence simple de la suite des fonctions caractéristiques des v.a. X_n vers une fonction continue en 0, alors cette fonction limite est la fonction caractéristique d'une v.a. X et il y a convergence en loi des X_n vers X), H. CRAMÉR[50] et E. LUKACS[51].

LAPLACE reprend ses résultats dans son ouvrage *Théorie analytique des probabilités* de 1812, en particulier dans le Livre I, *Calcul des fonctions génératrices* et dans le chapitre IV du livre 2, *De la probabilité des erreurs des résultats moyens d'un grand nombre d'observations, et des résultats moyens les plus avantageux*. Comme SIMPSON 50 ans plus tôt, LAPLACE commence par étudier le cas d'erreurs possibles de valeurs entières $-n, -(n-1),..., -1, 0, 1,..., n$ avec des probabilités égales. Pour s observations indépendantes, la probabilité que la somme des erreurs soit k est le terme en e^{ikw} dans $\left(\frac{1}{2n+1} \sum_{x=-n}^{n} e^{ixw}\right)^s$. Il détermine alors le *coefficient de Fourier* d'ordre k de cette série trigonométrique par :

---

[49] LÉVY, Paul (1937) : *Théorie de l'addition des variables aléatoires*. Gauthier-Villars, Paris.

[50] CRAMER, Harald (1937) : *Random variables and probability distributions*, Cambridge University Press (2ème éd., 1961).

[51] LUKACS, E. (1970) : *Characteristic Functions*. C. Griffin, London.

$$\frac{1}{2\pi}\left(\frac{1}{2n+1}\right)^s \int_{-\pi}^{\pi} dw\, \mathrm{e}^{-ikw}\left(\sum_{x=-n}^{n} e^{ixw}\right)^s ;$$

il calcule la probabilité que la somme des erreurs soit dans un intervalle $[-a\sqrt{s}\,;a\sqrt{s}]$ en assimilant la somme des probabilités individuelles à une intégrale et en développant les exponentielles en séries en gardant uniquement le premier terme en e^{-w^2}. Il considère ensuite le cas d'une loi de facilité quelconque, prend une intégrale en x au lieu d'une somme (i.e. la fonction caractéristique), développe en série le logarithme de la fonction à intégrer et ne conserve que le terme principal pour obtenir la probabilité d'un intervalle de la loi des erreurs ; il fait ici l'inversion de la fonction caractéristique et utilise de façon implicite la propriété de continuité démontrée par P. LÉVY plus d'un siècle après, résultat que certains auteurs appellent *théorème de Lindeberg-Lévy*.

LAPLACE revient à plusieurs reprises dans ses différents mémoires et sa *Théorie analytique des probabilités* sur la méthode des moindres écarts, due à BOSCOVICH, qu'il nomme *méthode de situation*, et il montre que *la médiane est le meilleur estimateur avec cette méthode*, ce qui rend nécessaire l'utilisation de la première loi de Laplace, à mettre en parallèle avec la démonstration de GAUSS que *la moyenne est le meilleur estimateur avec la méthode des moindres carrés* rend nécessaire la loi normale. LAPLACE a aussi montré (1811, *op. cit.*) que la méthode des moindres carrés est la plus avantageuse en ce sens qu'elle minimise l'erreur moyenne.

LAPLACE a ensuite utilisé son approche asymptotique dans plusieurs applications, en particulier la géodésie[52]. Son résultat, appelé maintenant Théorème-Limite Central est fondamental en statistique car, comme LAPLACE l'avait bien remarqué et insistait sur ce point, pour des observations nombreuses et sous des conditions assez larges[53], la loi de la moyenne est pratiquement une loi normale et ne dépend pas de la loi de distribution des observations individuelles. LAPLACE se fiait à son résultat théorique pour les applications, BESSEL (1818)[54] a fait une vérification expérimentale de la loi des erreurs sur des observations nombreuses ; ses valeurs s'ajustaient bien à une distribution normale.

---

[52] LAPLACE, P. S. (1818) : Application du calcul des probabilités aux opérations géodésiques, 2ème supplément de la *Théorie analytique des probabilités*. Il y a aussi un 3ème supplément sur le même sujet.

[53] La recherche de conditions de moins en moins restrictives et l'extension du Théorème-Limite Central à des espaces plus généraux a démarré à la fin du XIXe siècle et est devenue un champ entier de la théorie des probabilités et de la statistique.

[54] BESSEL, Friedrich (1818) : *Fundamenta Astronomiae*. Cf. MOLK, *Encyclopédie des sciences mathématiques*...

VII - La courbe en cloche et les sciences sociales au XIX^e^ siècle

Adolphe QUETELET avait participé activement au projet de créer un observatoire à Bruxelles et a été envoyé en voyage d'étude à Paris en 1823 pour apprendre l'astronomie pratique et la gestion d'un observatoire (il a été le premier directeur de l'Observatoire Royal de Bruxelles).

A Paris il s'est intéressé au calcul des probabilités et à la statistique, probablement avec J. FOURIER [FOURIER a dirigé le bureau de statistique de la ville de Paris et du département de la Seine dans les années 1820]. De retour en Belgique, il s'est occupé de statistiques de population et projetait un recensement d'après la « *méthode de M. De Laplace* »[55], et a continué simultanément avec son poste de directeur de l'observatoire royal de Bruxelles.

Dans les années 1830, QUETELET[56] a conçu sa notion d'*homme moyen* comme l'équivalent du centre de gravité pour des mesures faites sur des hommes différents, puis il a renversé son point de vue dix ans après ; les hommes (d'un âge, d'une race, d'un pays donnés) sont des exemplaires plus ou moins conformes de l'homme typique de cet âge, race, pays, et les mesures faites sur des hommes différents sont considérées comme des mesures différentes d'un même individu, l'*homme typique* (de l'âge, race, pays), d'une façon similaire à l'astronomie où il est fait des mesures différentes d'un même objet céleste.

En ce sens la *théorie des erreurs* de l'astronomie s'applique aussi aux mesures de caractéristiques humaines et la courbe des erreurs, en forme de cloche, doit aussi intervenir dans ce domaine. QUETELET considérait la mesure comme le résultat final d'un modèle mécaniste – le schéma d'urne binomial – somme de nombreuses petites causes accidentelles indépendantes. Il avait auparavant vérifié sur une distribution binomiale correspondant à 1 000 tirages l'ajustement à une courbe normale qu'il nommait loi ou courbe de possibilité, et il a appliqué une méthode analogue sur des mesures de tour de poitrine de soldats écossais et a trouvé que l'ajustement était très bon[57].

Comme pour ce qui se passait en astronomie, QUETELET considérait que les phénomènes en science morale (ou physique sociale) suivent aussi la loi des erreurs. Pour des observations, QUETELET utilisait l'ajustement à une courbe des erreurs comme un test sur l'homogénéité de la population. Si l'histogramme construit sur les observations n'avait pas la forme d'une courbe en cloche, c'était le

---

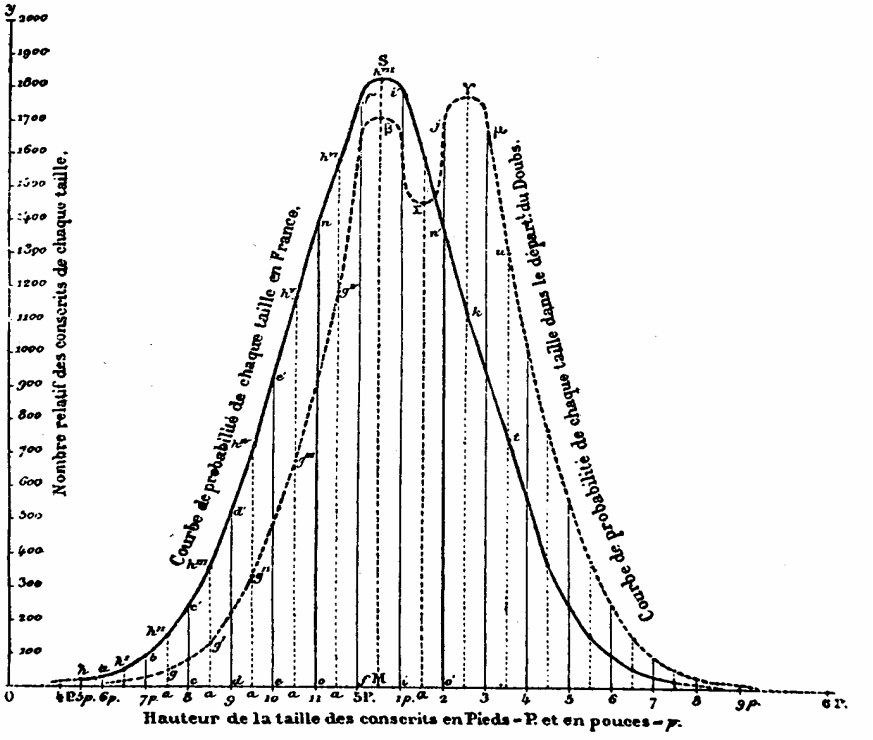
[55] QUETELET, Adolphe (1827) : Recherches sur la population, les naissances, les décès, ... dans le royaume des Pays-Bas, *Nouveaux mémoires de l'Académie des sciences et belles-lettres de Bruxelles*.

[56] QUETELET, Adolphe (1835) : *Sur l'homme et le développement de ses facultés, ou essai de physique sociale*, Paris.

[57] QUETELET, Adolphe (1846) : *Lettres à S.A.R. le Duc Régnant de Saxe-Cobourg et Gotha sur la théorie des probabilités,* ... Bruxelles.

signe d'une hétérogénéité dont il fallait chercher la cause et ensuite il pouvait ramener les observations à une courbe des erreurs.

Un exemple frappant de ceci a été donné par Adolphe BERTILLON[58] sur les tailles des conscrits du département du Doubs en 1863, cette courbe à double bosse étant expliquée par un mélange de types différents.



QUETELET a joué un rôle important par son incitation à la constitution de sociétés nationales de statistique (en particulier, la Royal Statistical Society en Angleterre) et la création du Congrès International de Statistique. Le premier Congrès s'est tenu à Bruxelles en 1853.

VIII - La courbe en cloche et l'école biométrique anglaise (fin XIX^e^ siècle et début XX^e^ siècle)

F. GALTON, dans son essai pour donner un fondement expérimental à la théorie de l'évolution de George DARWIN, a utilisé la méthodologie de QUETELET, et en particulier celle d'ajustement de données à une courbe des erreurs, appelée dans ce contexte loi des écarts. Pour QUETELET, s'il y a persistance des causes, alors il y aura une tendance pour que la moyenne soit stable et inversement une modification

[58] BERTILLON, Adolphe (1876) : Moyenne, *Dictionnaire encyclopédique des sciences médicales*, 2ème série, Paris

de la moyenne indiquera une variation des causes, une évolution. Dans le cas de nombreuses petites causes indépendantes et de même influence, la résultante suivra une loi des erreurs ; c'est le cas en particulier du schéma binomial que GALTON a illustré par son *quincunx* (ou planchette de Galton)[59]. Dans ses études sur la régression et la corrélation, GALTON a introduit la surface de corrélation qui correspond à une loi normale bivariée[60].

Les diverses expérimentations menées par GALTON en anthropométrie, le biologiste Walter WELDON, ami de Karl PEARSON, sur les crevettes, crabes... et d'autres, ont montré que les distributions de fréquence des mesures effectuées sur différents phénomènes ne suivaient pas toujours la courbe des erreurs. Avec son quincunx à deux étages, en remplissant de billes seulement deux compartiments intermédiaires, GALTON obtenait au bas de son appareil une distribution à deux bosses, mélange de deux distributions normales avec des moyennes différentes, du genre de celle que BERTILLON avait observé expérimentalement (voir ci-dessus).

Karl PEARSON a proposé un système de paramètres calculés à partir des moments d'ordre 1 à 4, pour caractériser la forme de ces distributions empiriques. La moyenne μ est un paramètre de position, ou tendance centrale ; la variance μ_2 est un paramètre de dispersion ; les moments centrés d'ordre 3 et 4 [si X est une variable aléatoire, $\mu_k = E((X-\mu)^k)$] vont servir à PEARSON pour définir des paramètres de forme : le coefficient d'asymétrie (skewness) $\beta_1 = \frac{\mu_3}{\mu_2^{3/2}}$, et le coefficient d'aplatissement (kurtosis) $\beta_2 = \frac{\mu_4}{\mu_2^2}$. En utilisant la notation σ pour l'écart-type (standard deviation) – notation et nom introduits par PEARSON en 1894 – ils s'écrivent $\beta_1 = \frac{\mu_3}{\sigma^3}$ et $\beta_2 = \frac{\mu_4}{\sigma^4}$; en particulier $\beta_2 = 3$ pour la courbe normale. D'autres mesures d'asymétrie et d'aplatissement ont été proposées, en particulier par R.A. FISHER, e.g. $\gamma_2 = \beta_2 - 3$. Ces paramètres sont encore largement utilisés en statistique exploratoire.

W. GOSSET (STUDENT) a illustré de façon humoristique dans un article[61] l'influence du coefficient d'aplatissement sur la forme de la courbe qui est mésokurtique pour $\beta_2 = 3$, cas de la loi normale, platykurtique pour $\beta_2 < 3$ (plus plate que la courbe normale) et leptokurtique pour $\beta_2 > 3$ (plus pointue que la courbe normale).

---

[59] Voir les articles *Expérimentation et simulation probabiliste* et *Du modèle à sa réalisation. La planche de Galton réalise-t-elle vraiment une distribution binomiale ?* dans ce même volume.

[60] GALTON, Francis (1889) : *Natural Inheritance*. London. Voir aussi PEARSON, Karl (1914-1930) : *The Life, Letters and Labours of Francis Galton* (3 vols.), Cambridge.

[61] STUDENT : Errors of routine analysis, *Biometrika* **19** (1927).

platykurtique

leptokurtique

Ensuite Karl PEARSON a proposé un système de distributions théoriques[62], c'est-à-dire de fonctions qui sont des densités de probabilité et vérifient l'équation différentielle $\frac{dy}{dx} = \frac{y(x+a)}{b_0 + b_1 x + b_2 x^2}$ (*) sur un certain domaine D de valeurs de x. Les valeurs des paramètres a, b_0, b_1, b_2 sont déterminées par les moments d'ordre 1 à 4 de la loi ayant pour densité la fonction y. Suivant les valeurs des paramètres a, b_0, b_1, b_2, PEARSON obtient différents types de distribution ; par exemple :

- si $b_0 \neq 0$, $b_1 = b_2 = 0$, les solutions de (*) donnent les lois normales $N(\mu, \sigma^2)$ avec $b_0 = -\sigma^2$ et $a = -\mu$;
- si $b_1 \neq 0$, $b_2 = 0$, les lois exponentielles et Gamma ;
- si $b_0 \neq 0$, $b_2 \neq 0$ tels que le trinôme $b_0 + b_1 x + b_2 x^2$ est irréductible sur $\mathbb{R}$, les fonctions puissance de ce trinôme, dont les lois de Cauchy et de Student ; si ce trinôme est réductible sur $\mathbb{R}$, les lois Bêta ; etc.

De plus, K. PEARSON a bien vu le rôle important des fonctions eulériennes en statistique, d'abord directement comme densités de probabilité de variables aléatoires, e.g. dans des processus à temps d'attente (pour les lois Gamma), et ensuite en relation avec le calcul numérique de diverses lois utilisées pour des tests statistiques : χ^2, Fisher, etc.[63]

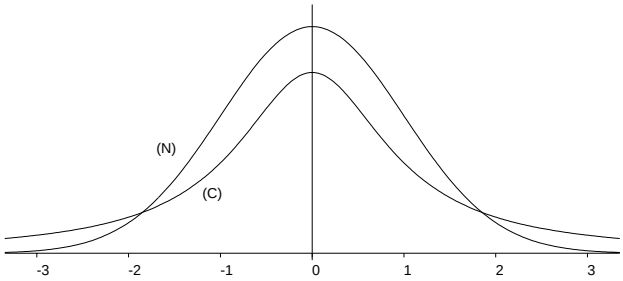
---

[62] PEARSON, Karl : Contributions to the mathematical theory of evolution, *Philosophical Transactions of the Royal Society of London* (1893-95) puis : Mathematical contributions to the theory of evolution, *Philosophical Transactions of the Royal Society of London* (1896-1905).

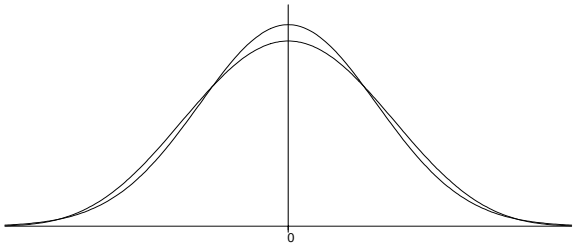
[63] PEARSON, Karl (1934) : *Table of the Incomplete Beta-function*, 2ème éd., 1968, Pearson, E. S. et Johnson, N. L. eds., Biometrika Trustees.

Les lois de Cauchy et de Student ont aussi des courbes de densité en forme de cloche, de même que sur un intervalle borné les lois Bêta symétriques, par exemple sur [0 , 1] la loi de densité $f(x) = Kx^{\lambda-1}(1-x)^{\lambda-1}$ pour $\lambda > 0$ (K étant une constante strictement positive telle que $\int f\,dx = 1$). Ainsi une représentation d'une distribution par une courbe en cloche, surtout s'il n'y a pas d'indication des unités, ne permet pas de savoir s'il s'agit d'une loi normale ou d'un autre type de loi de probabilité.

Comparaison des courbes de la loi normale (N) et de la loi de Cauchy (C)



Devinette : de ces deux graphes, quel est celui de la loi normale, de la loi Bêta (8,8) ?



Appendice :

Remarques sur les dénominations de théorèmes

Pour les théories, résultats et théorèmes qui nous semblent aujourd'hui importants, que ce soit en mathématiques, physique, ... ou sciences humaines, il est habituel de leur attribuer un nom, en particulier dans l'enseignement, pour pouvoir y référer commodément, et l'idéal serait que ce soit le nom des inventeurs ou initiateurs. Néanmoins, cette juste attribution d'un résultat à son ou ses créateurs se heurte à trois écueils :

(1) La publication des Mémoires des Académies, des livres et autres ouvrages demandait plusieurs mois et souvent quelques années, ensuite il y avait un délai parfois assez long pour la diffusion, de sorte que des mathématiciens pouvaient s'occuper des mêmes problèmes à la même époque et trouver indépendamment

des résultats relativement semblables. Pour cette raison, de temps en temps s'élevaient des querelles de priorité concernant un concept ou une théorie ; citons e.g. la querelle entre LEIBNIZ et NEWTON sur l'invention du calcul différentiel et intégral, et entre MONTMORT et DE MOIVRE au sujet de l'invention de certaines méthodes de résolution en calcul des probabilités. D'où aussi la difficulté d'attribution d'un nom, quelques décennies plus tard, pour un concept ou une théorie qui arrivait à une forme à peu près définitive après plus d'un demi-siècle de recherches de divers auteurs. De plus, certains travaux qui plus tard paraissent importants sont passés inaperçus au moment de leur publication.

(2) Jusqu'au premier tiers du XIX^e^ siècle et même jusqu'à aujourd'hui dans les manuels d'enseignement secondaire ou supérieur, l'art de la référence à des auteurs plus anciens, dont les résultats sont utilisés, était pratiquée de façon plus ou moins sporadique. Par exemple, LAPLACE fait très peu de références à des travaux antérieurs, sauf dans la notice historique de son *Essai philosophique sur les probabilités*.[64] Cependant, des revues scientifiques et les encyclopédies en France et dans les grands pays d'Europe (e.g. la grande *Encyclopédie* en France au milieu du XVIII^e^ siècle) passent périodiquement en revue les résultats qui semblent importants.

(3) Le poids de la tradition[65] et le fait que peu de mathématiciens, qu'ils écrivent des articles ou ouvrages de recherche ou qu'ils soient auteurs de manuels scolaires ou universitaires, s'intéressent à la justesse de l'attribution d'un théorème à un auteur antérieur ; ils sont pourtant très pointilleux sur la justesse des hypothèses ou conditions et la démonstration de ce même théorème[66].

Parmi les concepts étudiés dans cet article, le cas de la loi de probabilité (la loi des erreurs) qui est probablement la plus importante en théorie probabiliste et la plus utilisée en statistique, est symptomatique. Sur le point (1), on trouve pour l'approximation numérique de la probabilité d'une valeur et d'un intervalle : A. DE MOIVRE (1733) ; pour l'approximation par une intégrale de la probabilité d'un intervalle : Daniel BERNOULLI (1771, publié en 1773), et P. S. LAPLACE (1774) ; comme intégrale intervenant lors de l'approximation de fonctions dépendant de très grands nombres : P. S. LAPLACE (1785-86) ; comme loi de probabilité qui rend le plus probable le résultat donné par la méthode des moindres carrés : C. F. GAUSS (1809) et R. ADRAIN (1809) ; comme loi de probabilité limite de la somme d'un grand nombre d'erreurs élémentaires indépendantes ayant des variances du même ordre de grandeur : P. S. LAPLACE (1810-11).

---

[64] Il faut néanmoins noter la qualité à tous points de vue du livre de W. FELLER : *An Introduction to Probability Theory and its Applications*, (1950, 3^ème^ éd. 1968).

[65] Initiée probablement par des ouvrages d'enseignement.

[66] J'ai trouvé deux manuels d'enseignement universitaire de la fin du siècle dernier (1968, 1987 dont je tairai le nom des auteurs) qui appellent *intégrale de Gauss* le résultat $\int_0^\infty e^{-t^2}\,dt = \frac{\sqrt{\pi}}{2}$!

Si l'on considère que l'étude du comportement asymptotique de systèmes jouent un rôle moteur primordial depuis le milieu du XIX^e^ siècle en recherche mathématique et physique, dans la théorie des probabilités, la théorie ergodique, la théorie du potentiel, etc., et joue un rôle *central* en statistique, il est tout à fait justifié que cette loi soit dénommée *loi de Laplace*, ou *deuxième loi de Laplace* eu égard à la loi exponentielle bilatère nommée *première loi de Laplace* que LAPLACE a explicitement considérée comme une loi de facilité (probabilité) en 1774. Le terme *loi de Laplace-Gauss* pourrait en partie convenir aux partisans de l'exactitude, mathématique aussi bien qu'historique. Quant au poids de la *tradition* et la difficulté pour revenir sur une attribution incertaine, indiquée en (3), pour l'appellation de *loi de Gauss*, initiée probablement par J. BERTRAND, en particulier dans son livre[67] où il écrit dans la Préface (p. xxxiv) :

« La loi que doivent suivre, d'après une ingénieuse théorie, et que suivent à très peu près, quand elles sont nombreuses, les erreurs corrigées de toute inclination fixe, a été proposée par Gauss. ... »,

et plus loin dans le « *chapitre VIII - Loi des erreurs d'observation* » il écrit :

« Euler, Bernoulli, Lagrange et Laplace ont fait des hypothèses démenties par les faits et mal justifiées par des preuves sans vraisemblance. Gauss, plus heureux, a déduit d'un raisonnement fort simple une loi que la démonstration laisserait douteuse, mais que les conséquences justifient. ».

Karl PEARSON avait proposé en 1893 le nom de *loi normale* pour surmonter cette querelle d'attribution, et ce nom est maintenant très utilisé. D'un autre côté, malgré les efforts de certains auteurs de l'école probabiliste et statistique française depuis la fin du XIX^e^ siècle pour populariser le nom de *seconde loi de Laplace* (en particulier, P. LEVY[68] écrit en 1966 : « *Les lettres… désigneront toujours des variables laplaciennes réduites (rappelons que nous suivons M. Fréchet qui a proposé d'appeler loi de Laplace, la loi autrefois désignée sous le nom de loi de Gauss)* »), ou celui plus neutre de *loi de Laplace-Gauss* (par exemple FRÉCHET[69], SAPORTA, FOATA et FUCHS[70]), dans toute la littérature internationale et presque toute la littérature française, le terme consacré est maintenant *gaussienne* pour qualifier la loi ou distribution et *gaussien* pour un processus correspondant.

---

[67] BERTRAND, Joseph (1822-1900) : *Calcul des probabilités*, Gauthier-Villars, Paris, 1889 (2ème éd. 1907).

[68] LEVY, P., Fonction Browniennes dans l'Espace Euclidien et dans l'Espace de Hilbert dans *Festschrist for J. Neyman. Research Papers in Statistics*. Ed. F. N. David, John Wiley & Sons, London-New York, 1966. Citation de la p. 190.

[69] FRÉCHET, Maurice : *Généralités sur les Probabilités. Variables aléatoires*. Gauthier-Villars, Paris, 1937.

[70] Voir dans les références.

Références

DROESBEKE, J. J. et TASSI, Ph., *Histoire de la statistique*, Presses Universitaires de France, Collection Que Sais-je ?, 1990.

Encyclopédie des Sciences Mathématiques pures et appliquées, édité par Jules Molk, Gauthier-Villars et Teubner, Paris, 1904-1916. Réédité, J. Gabay, 1992. Tome I, Arithmétique et Algèbr. Vol. 4, Calcul des probabilités. Théorie des erreurs. Applications diverses.

KENDALL, M. G. and PLACKET, R. L. éds., *Studies in the history of statistics and probability*, vol. 2, C. Griffin & Co, Londres, 1977.

LACOMBE, H. et COSTABEL, P., éds., *La figure de la Terre du XVIII^e siècle à l'ère spatiale*, Gauthier-Villars, 1988.

LAPLACE, Pierre Simon de, *Oeuvres Complètes*, 14 tomes, Gauthier-Villars, Paris, de 1878 à 1912.

LAPLACE, Pierre Simon de, *Théorie analytique des probabilités* (1ère édition 1812, 2ème éd. 1814, 3ème édition 1820, *Oeuvres Complètes,* t. VII, 1886). Réédition J. Gabay (2 vols., 1995)

LAPLACE, Pierre Simon de, *Essai philosophique sur les probabilités* (1814, 5ème édition, 1825), préface de René THOM, postface de B. BRU, Editions Bourgois, 1986.

PEARSON, E. S. and KENDALL, M. G. eds., *Studies in the history of statistics and probability*, vol. 1, C., Griffin & Co, Londres, 1970.

PEARSON, K., *The History of Statistics in the 17th & 18th Centuries*, ed. E.-S. Pearson, Griffin & Co, Londres, 1978.

STIGLER, Stephen M., *The history of statistics*, *The Measurement of Incertainty before 1900*, Harvard University Press, 1986.

TODHUNTER, Isaac, *A History of the Mathematical Theory of Probability*, Cambridge, 1865. Rééd. Chelsea, New York, 1965.

Pour un traitement moderne des concepts abordés ici, voir par exemple :

FOATA, Dominique et FUCHS, Aimé, *Calcul des probabilités*, Masson, Paris, 1996, 2ème édition, 1998.

MÉTIVIER, M., *Notions fondamentales de la théorie des probabilités*, Dunod, Paris, 2ème édition, 1972.

Et pour la statistique,

SAPORTA, Gilbert, *Probabilités, analyse des données et statistique*, éd. Technip, Paris, 1990.

Introduction aux tests d'hypothèses, exemples

Michel HENRY, Annette CORPART

I - Exemple introductif : les faiseurs de pluie[1]

1 - La question

Des relevés effectués pendant de nombreuses années ont permis d'établir que le niveau naturel des pluies dans la Beauce, en millimètres par an, fluctue autour d'une valeur moyenne μ_0 qui, en 1950, était évaluée à 600 mm. Cette variation due aux aléas climatiques avait été mesurée par l'écart-type $\sigma_0 = 100$ mm. Remarquons dès maintenant la nécessité d'une statistique de référence.

Des entrepreneurs, surnommés *faiseurs de pluie*, prétendaient pouvoir augmenter de 50 mm le niveau moyen annuel de pluie, ceci par insémination des nuages par de l'iodure d'argent. Leur procédé fut mis à l'essai entre 1951 et 1959 et on releva les hauteurs de pluie suivantes :

année	1951	1952	1953	1954	1955	1956	1957	1958	1959
mm	510	614	780	512	501	534	603	788	650

Que pouvait-on en conclure ? Deux thèses s'affrontaient :

- ou bien l'insémination est sans effet,
- ou bien elle augmente réellement le niveau moyen des pluies d'au moins 50 mm.

Les agriculteurs étaient très intéressés par une augmentation réelle du niveau des précipitations, les 50 mm supplémentaires auraient notablement amélioré les rendements de leurs cultures, mais ils voulaient éviter une dépense non négligeable si les prétentions des faiseurs de pluie s'avéraient être des affabulations. Ils n'auraient accepté de s'engager dans ces dépenses qu'avec les meilleures garanties de réussite. Ils étaient typiquement devant un problème de test d'hypothèses, c'est-à-dire de prendre une décision entre deux hypothèses retenues pour l'étude (bien d'autres hypothèses auraient pu être formulées).

[1] Petite fable statistique où la réalité se mêle à la fiction, exemple et données numériques empruntées à G. SAPORTA : *Probabilités, analyse des données et statistique*, éd. Technip, 1992.

2 - Principe d'un test d'hypothèses simples

Soit μ la hauteur moyenne du niveau des précipitations après insémination des nuages. Ce paramètre est l'inconnue du problème.

Formulons deux hypothèses :

$H_0 : \mu = \mu_0$ ($\mu_0 = 600$ mm)

$H_1 : \mu = \mu_1$ ($\mu_1 = 650$ mm)

Pour simplifier cette introduction, nous avons posé $H_1 : \mu = \mu_1$ (hypothèse simple), plutôt que $H_1 : \mu > \mu_1$ (hypothèse multiple ou composite). Les agriculteurs auraient préféré s'en tenir à l'hypothèse H_0 (appelée *l'hypothèse nulle*), selon laquelle le procédé des faiseurs de pluie était sans effet sur les précipitations, à moins que les faits observés ne la contredisent nettement. Ils n'étaient décidés à abandonner H_0 pour l'hypothèse H_1 (*hypothèse alternative*) qu'en présence de faits expérimentaux traduisant une éventualité improbable, lorsque l'on suppose que H_0 est vraie. H_0 est l'hypothèse de prudence pour le client du test, H_1 est l'hypothèse dans laquelle on prend une décision risquée, celle de changer de pratique, ce qui représente un investissement qui suppose que cela soit justifié.

On peut donner un sens probabiliste à cette notion de *risque*, en considérant que le test conduit à une décision qui dépend de l'aléa de l'échantillon observé à partir duquel le test va conclure. Le contexte présent (H_0 est l'hypothèse de prudence) conduit à privilégier le risque de se tromper quand le test fait opter pour l'efficacité du procédé des faiseurs de pluie (événement que nous noterons H_1*), alors qu'il était sans effet (H_0 vraie). Ce risque est appelé le *risque de première espèce*, on l'interprète comme une probabilité : celle de « décider H_1 alors que H_0 est vraie », que nous noterons $P_{H_0}(H_1^*)$.

Le principe de la démarche dans un test d'hypothèses est de se fixer une valeur α, dite *seuil de signification* du test (on dit aussi que le test est de *niveau* $1 - \alpha$), et d'imposer que le risque de première espèce soit majoré par α. Pour la suite, prenons la valeur $\alpha = 0{,}05$ (une valeur traditionnelle dans un tel contexte) comme valeur maximale du risque qu'on accepte de prendre en rejetant à tort l'hypothèse H_0. Dans la pratique, on s'efforce de choisir H_1 de telle sorte que $P_{H_0}(H_1^*)$ soit égal à α, ou s'en rapproche le plus possible, ce qui a l'avantage, comme nous le verrons, de minimiser le risque dit de *seconde espèce*, égal à la probabilité de « conserver l'hypothèse H_0 alors que H_1 est vraie » : $P_{H_1}(H_0^*)$.

Les agriculteurs étaient prêts à accepter H_1 si le résultat de l'expérimentation réalisait une éventualité peu probable (de probabilité inférieure à 0,05) dans l'hypothèse H_0. Dans leur démarche, faisant confiance au hasard, ils admettaient implicitement que des événements rares ne sauraient se produire lors de l'expérimentation, et qu'en l'occurrence, la sagesse serait de remettre en cause le bien-fondé de l'hypothèse de travail, H_0. Ce faisant, ils assumaient le risque de se

tromper dans au plus 5 cas sur 100, cas où précisément les événements *rares* se produisent quand même. En principe, la valeur choisie pour α devrait faire intervenir le coût du procédé, face à l'accroissement espéré de productivité, de telle sorte que l'adoption du procédé des faiseurs de pluie apparaisse comme rentable, compte tenu du risque de première espèce accepté que l'échantillon observé induise inutilement cette décision.

3 - La méthode de décision

Puisqu'il s'agit de *tester* la hauteur moyenne des précipitations annuelles μ, il est naturel d'utiliser l'estimateur le plus efficace, c'est-à-dire la moyenne, notée $\overline{x}$, des hauteurs de précipitations observées sur les 9 années expérimentées.

Dans notre exemple $\mu_0 < \mu_1$ et si H_0 est vraie, les grandes valeurs de $\overline{x}$ sont plus improbables que sous l'hypothèse H_1. On se donne la *règle de décision* suivante :

- si $\overline{x}$ est trop grand, supérieur à une *valeur critique* μ_c qui, dans l'hypothèse H_0 n'a que 5 chances sur 100 d'être dépassée, on optera pour H_1 avec par conséquent la probabilité $P_{H_0}(H_1^*) \leq \alpha = 0{,}05$ de se tromper ;
- si $\overline{x} \leq \mu_c$, on conservera H_0 faute de preuves suffisantes. Cela ne signifie en rien que H_0 soit effectivement vraie.

II - Fondements probabilistes d'un test d'hypothèses

1 - Modèle probabiliste pour le test d'hypothèses proposé

Formalisons le problème. La hauteur de pluie de l'année i ($1 \leq i \leq 9$) est représentée par une variable aléatoire X_i. Pour de nombreuses causes hasardeuses, cette grandeur fluctue de part et d'autre d'une moyenne μ avec une dispersion d'écart-type σ. Pour représenter ce phénomène météorologique, prenons un modèle gaussien[2]. Bien que la hauteur de pluie soit une variable naturellement positive, on suppose sans inconvénient appréciable que les X_i suivent une même loi normale $\mathcal{N}(\mu\,,\sigma)$. En effet, la probabilité de $]-\infty\,;0[$, est négligeable, aussi bien pour $\mathcal{N}(\mu_0\,,\sigma)$ que pour $\mathcal{N}(\mu_1\,,\sigma)$, prenant pour valeurs respectives $\Phi(-6)$ et $\Phi(-6{,}5)$, où Φ est la fonction de répartition de la loi $\mathcal{N}(0\,,1)$.

On a donc pour tout i, $E(X_i) = \mu$ et $Var(X_i) = \sigma^2$. Dans la suite, on prendra $\sigma = \sigma_0 = 100$ mm, sous H_0 ainsi que sous H_1, en considérant qu'il n'y a pas de raison que le procédé des faiseurs de pluie modifie notablement la dispersion du niveau des précipitations autour de sa moyenne.

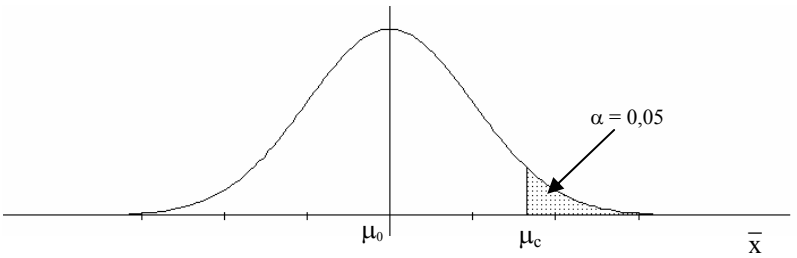
---

[2] Cf. l'article *Phénomènes gaussiens et loi normale* dans ce même volume. On oublie souvent de dire que la conclusion du test est tributaire de ce choix de modèle. Il en va de même pour l'hypothèse qui suit concernant la valeur de σ.

On considère aussi que les X_i sont indépendantes (pas d'influence d'une année sur l'autre), elles constituent donc un échantillon $X = (X_1, \ldots, X_9)$ de taille $n = 9$.

Soit $\overline{X}$ la moyenne arithmétique des X_i. On sait que, sous ces hypothèses, $\overline{X}$ suit la loi normale[3] $\mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. On a $E(\overline{X}) = \mu$ et $\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$. La valeur prise par $\overline{X}$ sur l'échantillon donné est notée $\overline{x}$.

La méthode de décision introduit une valeur critique μ_c, telle que si $\overline{x} > \mu_c$, on rejette l'hypothèse H_0 pour lui préférer H_1. La valeur optimale de μ_c est donc donnée par la condition[4] $P_{H_0}(\overline{X} > \mu_c) = \alpha$.



Repérage de la valeur critique μ_c à partir de la courbe représentative de la densité $f_{\overline{x}}$.

Sous l'hypothèse H_0, $\mu = \mu_0 (= 600)$ et $\text{Var}(\overline{X}) = \left(\frac{100}{3}\right)^2$; la loi de $\overline{X}$ est donc entièrement connue et on peut déterminer numériquement μ_c à partir de la fonction de répartition de la loi normale : la variable $U = \frac{\overline{X} - 600}{100} \times 3$ est normale centrée réduite. La condition critique $P(\overline{X} > \mu_c) = 0{,}05$ s'écrit $P\left(U > \frac{\mu_c - 600}{100} \times 3\right) = 0{,}05$, et on trouve dans la table $\frac{\mu_c - 600}{100} \times 3 \approx 1{,}65$, d'où $\mu_c \approx 655$ mm.

$\overline{x}$ étant la moyenne observée, on en tire alors la *règle de décision* :

- si $\overline{x} > 655$ mm, on rejette H_0 et on accepte H_1 avec un risque inférieur à $\alpha = 0{,}05$ de se tromper ;
- si $\overline{x} \leq 655$ mm, on conserve H_0, sans être certain qu'elle est vraie.

---

[3] Cf. l'article cité, § III.

[4] μ_c est donc le quantile d'ordre α de la loi P_{H_0} de $\overline{X}$. Dans notre cas, cette loi étant continue, l'égalité $P_{H_0}(\overline{X} > \mu_c) = \alpha$ peut être réalisée.

L'ensemble des valeurs de $\overline{X}$ qui conduisent à rejeter H_0 s'appelle la *région critique* ou *région de rejet* de H_0 ; dans notre exemple, c'est l'intervalle $]655 , +\infty[$.

Son complémentaire est la *région d'acceptation* de H_0, ici : $]-\infty , 655]$.

Dans l'exemple, on a $\overline{x} = 610{,}2$ mm. Les agriculteurs ont donc conservé H_0, doutant de l'affirmation des faiseurs de pluie et concluant que la légère augmentation moyenne des précipitations observées durant l'expérimentation était due au hasard et non à l'iodure d'argent.

2 - Risques, hypothèses et certitudes

L'observation d'un échantillon n'apporte jamais de certitude quant à la population. Rien ne dit ici que conserver H_0 mette à l'abri de se tromper : les faiseurs de pluie avaient peut-être raison, mais on ne s'en est pas aperçu. En fait, il y avait deux manières de se tromper :

- Croire les faiseurs de pluie alors qu'ils n'étaient pour rien dans le résultat obtenu. C'est l'erreur de première espèce associée au risque de première espèce $P_{H_0}(H_1^*)$. Formellement, cela revient donc à accepter H_1 alors que H_0 est vraie. Rappelons que par construction, ce risque est majoré par le seuil de signification α du test : $P_{H_0}(H_1^*) \leq \alpha$.
- Ne pas croire les faiseurs de pluie alors que leur méthode est bonne et que seul le hasard (malencontreux pour eux) intervenant sur un petit nombre d'observations, a donné des résultats insuffisants pour convaincre les agriculteurs. C'est l'erreur de seconde espèce : conserver H_0 alors que H_1 est vraie. Cette décision dépend de l'aléa de l'échantillon observé, c'est donc un événement que nous notons H_0^*. Quand H_1 est vraie, la probabilité de conserver H_0 est $P_{H_1}(H_0^*)$, c'est le *risque de seconde espèce*.

Dans notre exemple, on peut calculer le risque de seconde espèce : sous l'hypothèse H_1, nous avons convenu que $\overline{X}$ suit la loi normale $\mathcal{N}\left(\mu_1 , \frac{\sigma}{\sqrt{n}}\right)$, avec $\mu_1 = 650$ mm et $\sigma = 100$ mm. On commet alors l'erreur de seconde espèce lorsque l'observation $\overline{x}$ est inférieure à $\mu_c = 655$ mm. La variable aléatoire $U' = \frac{\overline{X} - 650}{100} \times 3$ suit la loi normale $\mathcal{N}(0 , 1)$. On a :

$$P_{H_1}\left(\overline{X} < 655\right) = P\left(U' < \frac{655 - 650}{100} \times 3\right) = 0{,}56.$$

La seule certitude est que les agriculteurs ont pris un risque de se tromper important, supérieur à 0,5, en conservant H_0.

H_0, tout au long de cet exemple, a joué un rôle prépondérant : hypothèse de prudence, c'est celle que l'on conservera à moins d'être convaincu qu'il vaut mieux

l'abandonner pour H_1, l'hypothèse alternative. H_1 est l'hypothèse qui implique un investissement (par exemple acheter le procédé des faiseurs de pluie). Cela n'implique pas que H_1 soit le contraire de H_0 (ici, le contraire de H_0 serait $\mu \neq \mu_0$, hypothèse composite bilatérale).

Dans le processus de modélisation, le choix de H_0 est dicté par des mobiles assez variables :

(1) hypothèse de stabilité ; elle doit être solidement établie et n'a pas été jusque là contredite par l'expérience ;

(2) H_0 est une hypothèse à laquelle on tient particulièrement pour des raisons qui peuvent être subjectives ;

(3) H_0 est une hypothèse de prudence ; par exemple pour tester l'innocuité d'un vaccin, il vaut mieux partir d'une hypothèse défavorable au nouveau produit ;

(4) H_0 est une hypothèse facile à formuler (exemple $\mu = \mu_0$ de préférence à $\mu \neq \mu_0$) qui rend les calculs possibles.

En particulier, la loi de la variable de décision doit être entièrement connue sous l'hypothèse H_0.

3 - Région critique

La forme de la région critique $W = \left\{ \overline{X} > \mu_c \right\}$ est indiquée par la nature de H_1 (650 est plus grand que 600), mais la valeur de μ_c ne dépend que de H_0 et de α, comme on l'a vu en la calculant.

W est donc l'événement qui correspond à l'ensemble des valeurs prises par la variable de décision qui conduisent à écarter H_0 au profit de H_1.

– La région critique optimale W est celle qui maximise le risque de première espèce $P_{H_0}(W)$ tout en respectant la condition : $P_{H_0}(W) \leq \alpha$.

– La probabilité $P_{H_1}(W)$ est celle d'opter pour H_1 en ayant raison. On l'appelle la *puissance du test*, notée $1 - \beta$.

– La région d'acceptation de H_0 est donc le complémentaire W^c de la région critique, on a : $P_{H_0}(W^c) \geq 1 - \alpha$ et $\beta = P_{H_1}(W^c)$.

Ces valeurs sont résumées dans le tableau suivant donnant, dans la situation optimale, les probabilités des quatre situations possibles :

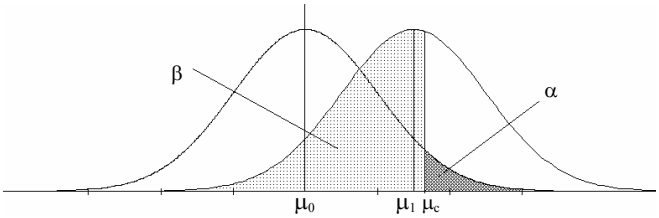
		Hypothèses à tester	
		H_0	**H_1**
Décision induite par le test	**H_0***	$1 - \alpha$	β
	H_1*	α	$1 - \beta$

Construire un test, c'est donc déterminer *a priori* la région critique W, ceci sans attendre de connaître le résultat de l'expérience.

4 - Puissance d'un test, détermination des risques

Dans le schéma ci-dessous, on a tracé les deux courbes de densité des lois de $\overline{X}$ sous les hypothèses respectives H_0 et H_1. Une fois μ_0 et α fixés, la valeur critique μ_c en découle, délimitant sous la première courbe une aire grisée la plus foncée représentant la valeur maximale α du risque de première espèce. La valeur μ_c fait alors apparaître sous la deuxième courbe le risque de seconde espèce β, comme le montre la figure : l'aire grisée la plus claire représente la probabilité que $\overline{X}$ soit inférieure à μ_c dans l'hypothèse H_1 où la moyenne μ est égale à μ_1.

Ce schéma montre aussi que les risques de première et seconde espèce varient en sens contraire quand, μ_0 et μ_1 étant fixés, on fait varier le seuil α.



Risques de première et seconde espèce

Lorsque μ_1 est proche de μ_0, on voit que pour un α fixé, le risque β peut être grand, on dit que *le test est peu puissant*. La *puissance* du test est la valeur $1 - \beta$, c'est la probabilité que le test conduise à choisir H_1 en ayant raison, cas où l'observation de l'échantillon donne $\overline{x} > \mu_c$.

Dans la pratique d'un test, on limite le risque de première espèce au seuil α**.** Il correspond au risque de se tromper en rejetant l'hypothèse préférée H_0. Ce faisant, on accepte un risque de deuxième espèce β, risque de se tromper en conservant H_0, qui peut être nettement plus important. Il vaut dans l'exemple $\beta = 56$ % quand $\mu_1 = 650$ mm. On pourrait penser minimiser ce risque en choisissant mieux l'hypothèse alternative H_1. Mais plus on prend μ_1 grand, plus β est petit et plus la probabilité d'opter pour H_1 avec raison est grande. Cependant une trop grande valeur pour μ_1 serait une exigence déraisonnable vis-à-vis du phénomène étudié et rencontrerait sans doute l'opposition des faiseurs de pluie.

Comme on l'a vu, la valeur du seuil de signification α est déterminée par des considérations concrètes, par exemple économiques. On évalue les coûts de commettre les erreurs de première et deuxième espèce ainsi que le gain que procure

un investissement à bon escient, et l'on prend α de telle sorte que la décision à prendre soit la plus rentable possible.

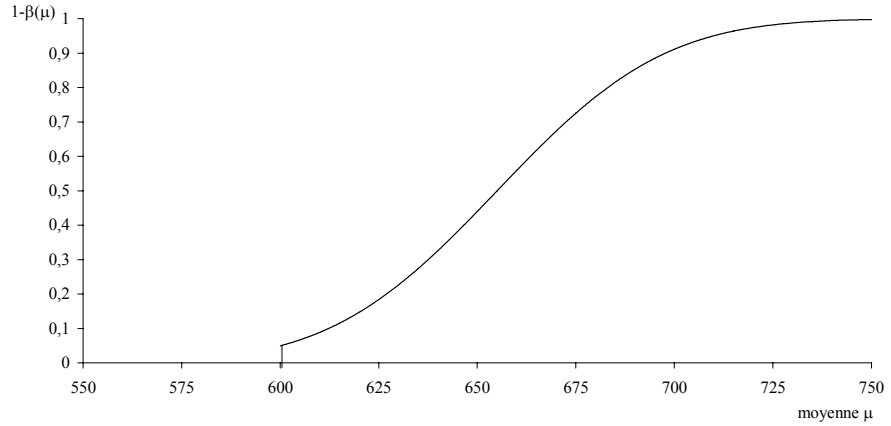
Le risque de se tromper en acceptant H_0 n'est pas une fonction simple de α : la valeur du risque de deuxième espèce β dépend de l'hypothèse alternative, de α et *de la loi de la variable de test* $\overline{X}$, donc de la taille n de l'échantillon prélevé pour l'observation. Si $\beta \to 0$ lorsque $n \to \infty$, on dit que le test est *convergent*.

5 - Cas d'une hypothèse alternative composite

Dans l'exemple des faiseurs de pluie, il semble plus naturel de proposer l'hypothèse alternative $H_1 : \mu > \mu_1$, les entrepreneurs garantissant une moyenne supérieure à un minimum μ_1. On dit que $\mu > \mu_1$ est une hypothèse unilatérale ($\mu \neq \mu_0$ serait une hypothèse bilatérale) composite ou multiple, car elle concerne un ensemble de valeurs du paramètre μ testé non réduit à un point (ici, l'ensemble continu $]\mu_1 , +\infty[$).

La région critique W, vu l'allure de l'hypothèse alternative, est de la forme $]\mu_c, +\infty]$, elle est entièrement définie par la condition optimale $P_{H_0}(\overline{X} > \mu_c) = \alpha$

Sous l'hypothèse H_1, à chaque valeur $\mu \in]\mu_1, +\infty[$ correspond une loi $\mathcal{N}\left(\mu , \frac{\sigma}{\sqrt{n}}\right)$ pour la variable de décision $\overline{X}$. H_1 est donc modélisée par cette famille de lois déterminant la famille de probabilités (P_μ) pour $\mu \in]\mu_1 , +\infty[$. A chaque valeur μ correspond donc un risque de seconde espèce $\beta(\mu) = P_\mu(W^c)$. On obtient ainsi la fonction puissance du test $1 - \beta(\mu)$, probabilités fonctions de μ (dont la « vraie » valeur reste inconnue) de rejeter H_0 en ayant raison. Dans notre exemple, prenant $\mu_1 = \mu_0 = 600$ mm, la courbe représentative de la fonction puissance a l'allure suivante :



Fonction puissance du test unilatéral $H_0 : \mu = 600$ mm contre $H_1 : \mu > 600$ mm

6 - Pratique d'un test d'hypothèses

Les tests d'hypothèses diffèrent en fonction de la forme des hypothèses envisagées et des variables de décision adoptées. Résumons les étapes dans la réalisation d'un test d'hypothèses. Un seuil α étant donné, on suit la démarche suivante :

1 - Modélisation de la situation : choix de H_0 et de H_1.

2 - Fabrication du test :

i - détermination de la variable de décision et de sa loi dans l'hypothèse H_0. Dans l'exemple : $\overline{X} \sim \mathcal{N}\left(\mu_0 , \frac{\sigma}{\sqrt{n}}\right)$.

ii - Allure de la région critique W en fonction de H. Dans l'exemple : W est de la forme $]\mu_c, +\infty[$.

iii - Détermination de W en fonction de H_0 et de α Dans l'exemple : $W =]655, +\infty[$.

iv - Calcul (éventuel) de la puissance du test, ce qui suppose de connaître la loi de la variable de décision sous l'hypothèse H_1. Dans l'exemple : $\overline{X} \sim \mathcal{N}\left(\mu , \frac{\sigma}{\sqrt{n}}\right)$.

3 - Mise en œuvre du test :

i - détermination de la valeur observée de la variable de décision au vu de l'échantillon. Dans l'exemple : $\overline{x} = 612$ mm.

ii - Conclusion : rejet ou acceptation de H_0. Dans l'exemple : les agriculteurs ont accepté H_0.

7 - Conclusion d'un test d'hypothèses

Les hypothèses H_0 et H_1 ne sont pas elles-mêmes probabilisables. Elles sont vraies ou fausses (mais on ne le sait pas). C'est le fait que l'observation de l'échantillon conduise à retenir l'une ou l'autre (événements H_0* ou H_1*) qui est aléatoire et qui peut donc être probabilisé.

Ainsi, *accepter l'hypothèse* H_0 (événement H_0*) signifie que l'échantillon observé donne à la variable de décision $\overline{X}$ une valeur $\overline{x}$ située dans la région d'acceptation. Dans l'exemple : $\overline{x} \leq \mu_c$.

Effectuer un test au seuil de 5 % signifie que, si H_0 est conforme à la réalité, ce test conduira à rejeter H_0 à tort, optant à tort pour H_1, pour 5 échantillons sur 100 en moyenne. Ainsi, 95 échantillons sur 100, en moyenne, conduiront à conserver H_0 avec raison.

Mais si c'est H_1 qui devrait être choisie, la probabilité de retenir H_0 à tort est le risque de seconde espèce β (dans l'exemple : 56 %), alors que la probabilité d'opter pour H_1 avec raison est $1 - \beta$ (ici 44 %).

La conclusion d'un test d'hypothèses ne peut donc en aucun cas être de la forme insensée : « La probabilité que H_0 soit vraie est 95 %. » !

III - Application : un test binomial (fantaisiste)

Un examinateur doit faire passer une épreuve type Q.C.M. à des étudiants. Ce Q.C.M. est constitué de 20 questions indépendantes. Pour chaque question, il y a trois réponses possibles dont une seule correcte. On considère ces 20 questions comme un échantillon de toutes celles qui pourraient être posées pour mettre en évidence les compétences des étudiants.

On suppose qu'il y a deux types d'étudiants (hypothèse très simplificatrice) :

- l'étudiant qui n'a pas travaillé et qui répond au hasard : il a alors une chance sur trois d'avoir une réponse juste.
- l'étudiant qui a travaillé : il a davantage de chances de donner une bonne réponse à chaque question mais le pourcentage de réussite est inconnu.

L'examinateur veut déterminer une valeur critique k_c telle que :

- si le nombre de réponses correctes est supérieur ou égal à k_c, l'étudiant est reçu,
- si le nombre de réponses correctes est strictement inférieur à k_c, l'étudiant est recalé.

Pour un étudiant donné, on considère la variable aléatoire X égale au nombre de ses réponses correctes aux 20 questions. X suit la loi binomiale $\mathcal{B}(20\,;p)$ où p est la probabilité que l'étudiant réponde juste à chacune des questions : si l'étudiant n'a pas travaillé, on a $p = \frac{1}{3}$.

Pour prendre une décision la plus juste possible (admettre ou recaler un étudiant), on fait un test, en considérant que la réponse de l'étudiant au Q.C.M. est un échantillon, pris au hasard dans la population, des réponses qu'il fournirait à tout Q.C.M. du même type.

1 - Choix des hypothèses :

Hypothèse nulle H_0 : l'étudiant n'a pas travaillé.

Hypothèse alternative H_1 : l'étudiant a travaillé.

(On notera le choix de l'hypothèse de prudence retenu par l'examinateur !)

Nous sommes donc en présence d'une hypothèse alternative H_1 unilatérale et composite, se traduisant par la condition $p > \frac{1}{3}$.

2 - Variable de décision :

On prend la variable X, égale au nombre de réponses justes, qui suit la loi binomiale $\mathcal{B}\left(20, \frac{1}{3}\right)$ sous l'hypothèse H_0.

3 - Allure de la région critique :

Sous H_1, l'étudiant a plus de chances de répondre juste à chaque question. La région critique est donc de la forme $\{X \geq k_c\}$.

A l'issue d'un test, quatre situations sont possibles :

1 : L'étudiant n'a pas travaillé et il est recalé

2 : L'étudiant a travaillé et il est reçu

3 : L'étudiant n'a pas travaillé et il est reçu

4 : L'étudiant a travaillé et il est recalé

On a donc deux types d'erreurs possibles correspondant aux situations 3 et 4 :

- l'erreur de première espèce : le test conduit à rejeter l'hypothèse H_0 (l'étudiant n'a pas travaillé) alors qu'elle est vraie (situation 3).
- Les erreurs de seconde espèce associées aux différentes valeurs possibles de la probabilité *p* qu'un étudiant qui a travaillé réponde juste à une question : si le test ne permet pas de rejeter H_0, on pense que l'étudiant n'a pas travaillé, alors que c'est faux. Conclusion injuste (situation 4).

On veut contrôler le risque de première espèce par un seuil α : on cherche à recaler la plupart des étudiants qui n'ont pas travaillé.

4 - Détermination de la région critique :

Les valeurs possibles pour X sont entières dans [0 ; 20]. La région critique est de la forme $[k_c ; 20]$.

Le risque de première espèce doit être majoré par le seuil α, la valeur critique k_c vérifie $\alpha \geq P_{H_0}(X \geq k_c)$. La table de la loi binomiale $\mathcal{B}\left(20, \frac{1}{3}\right)$ montre que :

- si on choisit $k_c = 10$, alors le seuil possible est limité : $\alpha \geq 9\ \%$.

- si on choisit $k_c = 15$, alors on peut avoir un seuil nettement meilleur : $\alpha \approx 0{,}02\ \%$.

On peut aussi fixer le seuil α (majorant le risque de se tromper en recevant un étudiant), on en tire alors la valeur critique k_c. Par exemple, pour $\alpha = 1\ \%$, $k_c = 12$.

5 - Puissance du test

Sous l'hypothèse composite H_1, pour chaque valeur de $p > 1/3$, X suit une loi binomiale $\mathcal{B}(20, p)$. À chaque $p > 1/3$ correspond un risque de seconde espèce : $\beta(p) = P_p(X < k_c)$. On obtient ainsi la fonction puissance du test $1 - \beta(p)$.

Prenons un exemple numérique pour fixer les idées. Supposons qu'un étudiant qui a travaillé double presque ses chances de répondre correctement aux questions. Soit $p = 0{,}6$ sa probabilité de réussite, supposée constante pour toutes les questions. Les tables et graphiques qui suivent donnent les probabilités binomiales pour $p = 1/3$ et $p = 0{,}6$. Ils montrent la tendance des probabilités binomiales à se répartir en forme de cloche (convergence vers une loi normale) et illustrent les variations en sens contraires des risques de première et seconde espèce (ici défini quand on fixe $p = 0{,}6$), lorsque l'on fait varier la valeur critique k_c. On obtient par exemple :

avec $k_c = 10$, $\alpha \geq 9\ \%$ et $\beta \approx 13\ \%$

avec $k_c = 12$, $\alpha \geq 1\ \%$ et $\beta \approx 40\ \%$

avec $k_c = 14$, $\alpha \geq 0{,}1\ \%$ et $\beta \approx 75\ \%$!

On peut arriver à réduire en même temps les deux risques en augmentant le nombre de questions.

Avec, par exemple, 40 questions et $k_c = 20$ (au moins la moitié des réponses correctes), on obtient $\alpha \geq P_{H_0}(X \geq 20) \approx 2\ \%$ et $\beta = P_{0,6}(X < 20) \approx 7\ \%$.

Tables de la loi binomiale

p = 1/3

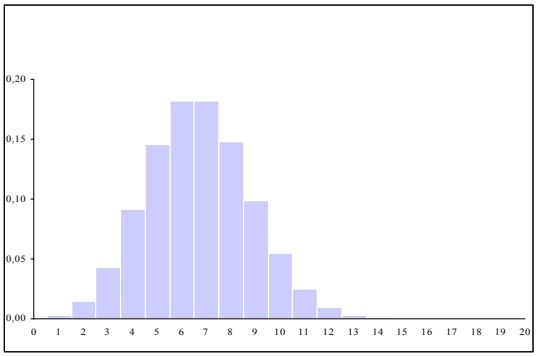
k	P(X = k)
0	0,000301
1	0,003007
2	0,014285
3	0,042854
4	0,091064
5	0,145703
6	0,182129
7	0,182129
8	0,147980
9	0,098653
10	0,054259

k	P(X = k)
11	0,024663
12	0,009249
13	0,002846
14	0,000711
15	0,000142
16	0,000022
17	0,000003
18	0,000000
19	0,000000
20	0,000000

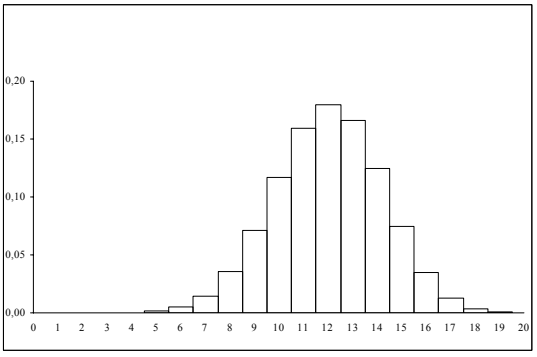
p = 0,6

k	P(X = k)
0	0,000000
1	0,000000
2	0,000005
3	0,000042
4	0,000270
5	0,001294
6	0,004854
7	0,014563
8	0,035497
9	0,070995
10	0,117142

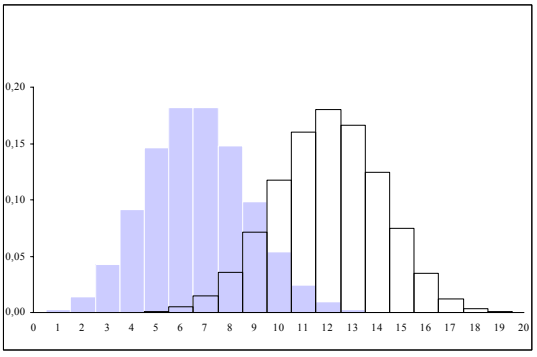
k	P(X = k)
11	0,159738
12	0,179706
13	0,165882
14	0,124412
15	0,074647
16	0,034991
17	0,012350
18	0,003087
19	0,000487
20	0,000037



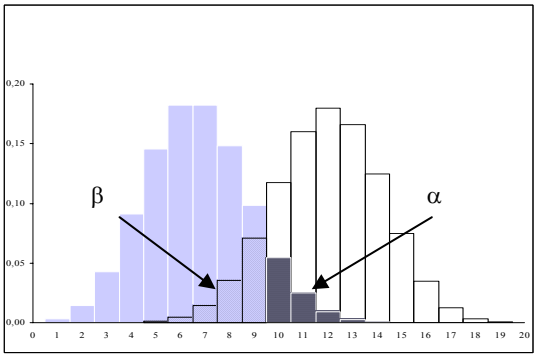
Loi binomiale $\mathcal{B}\left(20\,,\frac{1}{3}\right)$



Loi binomiale $\mathcal{B}(20\,,\,0{,}6)$



Comparaison



Seuil α et risque β pour la valeur critique $k_c = 10$

6 - Énoncé de la règle de décision :

Pour le test à 20 questions, désirant minimiser le nombre de reçus n'ayant pas travaillé, on choisit $k_c = 12$. Pour chaque étudiant :

- s'il a au moins 12 réponses correctes, alors on rejette l'hypothèse H_0, et on le déclare reçu.
- s'il a strictement moins de $k_c = 12$ réponses correctes, alors on n'a pas suffisamment de raisons de rejeter l'hypothèse H_0, et on le déclare recalé.

Dans le cas d'un étudiant qui n'a pas suffisamment travaillé, l'examinateur prend un risque de 1 % de l'avantager à tort.

Lorsqu'un étudiant a bien travaillé, doublant ainsi ses chances de bien répondre aux questions, l'examinateur prend alors un risque $\beta = 40\ \%$ de commettre une injustice !

Pour réduire ce risque inacceptable, il vaudrait mieux poser 40 questions et doubler le temps de travail de l'examinateur !

Tests d'adéquation à une loi de probabilité, pratique des tests du Khi-deux

Louis-Marie BONNEVAL[1] *et Michel HENRY*

Les programmes de terminale S et de terminale ES en vigueur à la rentrée 2002 comportent un paragraphe, identique dans les deux filières, intitulé *Simulation* et formulé comme suit :

« Etude d'un exemple traitant de l'adéquation de données expérimentales à une loi équirépartie ».

Le commentaire est le suivant :

« L'élève devra être capable de poser le problème de l'adéquation à une loi équirépartie et de se reporter à des résultats de simulation qu'on lui fournit. Le vocabulaire des tests[2] *(test d'hypothèses, hypothèse nulle, risque de première espèce) est hors programme ».*

C'est bien entendu aux tests du Khi-deux (appelés ainsi car ils font intervenir la loi d'une variable standard notée χ^2) qu'il est fait allusion.

L'introduction de cette notion avant le baccalauréat est une nouveauté, dont on ne peut contester l'intérêt : tester la validité d'un modèle est une démarche essentielle dans toute activité scientifique que beaucoup de nos élèves (y compris ceux qui se dirigeront vers les sciences humaines) auront à pratiquer.

Mais elle est délicate à mettre en œuvre, car elle suppose de la part des enseignants une vision claire autant de la théorie sous-jacente que de ses enjeux didactiques. Or beaucoup de collègues sont mal à l'aise avec la statistique inférentielle, n'ayant pas eu dans leur cursus la formation *adéquate*.

Pour contribuer à la compléter, ce chapitre propose une brève présentation des tests du Khi-deux et développe plus complètement le cas du contrôle de l'adéquation d'une distribution de fréquences à une loi de probabilité, en particulier équirépartie.

---

[1] L'essentiel de ce chapitre est repris d'un article de Louis-Marie BONNEVAL paru dans le n° 441 du bulletin de l'APMEP sous le titre : *Test d'équirépartition : qui a dit Khi-deux ?*

[2] Pour une initiation aux tests d'hypothèses, on pourra consulter CARNEC, H. et al., *Itinéraires en statistiques et probabilités*, Ellipses, 2000. Pour un approfondissement, voir KAUFFMANN, P., *Statistique, Information, Estimation, Tests*, Dunod, 1994.

I - Principe des tests d'adéquation et champs d'applications

Les valeurs observées d'un caractère *C* défini sur une population statistique *P* sont souvent réparties entre k modalités M_i qui constituent une partition du domaine de variations de ce caractère. C'est directement le cas quand *C* est un caractère qualitatif. Il en est de même quand *C* est quantitatif discret ou continu, l'étude de la répartition de ses valeurs possibles se limite alors à leur distribution entre un nombre fini de classes M_i dont l'explicitation constitue le modèle général dans lequel se situera l'étude.

A partir d'un échantillon prélevé dans la population *P,* le principe d'un test d'adéquation consiste à comparer la distribution des fréquences observées $(f_i)_{1\le i\le k}$, notée (f), des différentes modalités M_i du caractère C, avec une loi de probabilité *théorique* $(p_i)_{1\le i\le k}$, notée (p) par la suite. Cette loi de probabilité constitue un *sous-modèle probabiliste* censé[3] représenter les variations du caractère *C* dans la population.

Dans cette hypothèse, p_i serait la valeur de la probabilité que le caractère d'un élément pris au hasard dans la population *P* soit de modalité M_i. Tenant compte des fluctuations d'échantillonnage, les fréquences observées f_i seront considérées comme les valeurs prises sur l'échantillon par une famille de variables aléatoires F_i, notée (F) par la suite. Le test apprécie la proximité des deux familles (f) et (p), et, le cas échéant, permet de conclure à *l'adéquation de la distribution des fréquences observées à la loi théorique donnée*.

Pour pouvoir appliquer des résultats probabilistes puissants, on considère donc que l'échantillonnage est aléatoire (observation de n éléments pris au hasard dans la population, sans modifier les probabilités[4] des différentes modalités). On suppose aussi que les éléments prélevés le sont indépendamment les uns des autres.

D'un échantillon à l'autre, les fréquences observées fluctuent, reflétant les aléas des prélèvements. Elles respectent cependant la contrainte $\sum_{i=1}^{k} f_i = 1$. On dit que la distribution de fréquences (f) a $k - 1$ *degrés de liberté*.

On suppose donc que p_i est effectivement la probabilité de la modalité M_i. On sait d'expérience que quand la taille n de l'échantillon augmente, les fréquences observées f_i *tendent à se stabiliser*[5] vers les probabilités p_i. Les lois des F_i peuvent

---

[3] Comme pour toute modélisation, un tel modèle n'est qu'un reflet simplifié et hypothétique de la réalité. La question est d'en choisir un qui rende compte au mieux de cette réalité.

[4] La meilleure connaissance de ces probabilités est précisément l'objet du test du Khi-deux.

[5] Cette expression, qui traduisait l'introduction *fréquentiste* de la notion de probabilité dans les programmes de première de 1991, exprime naïvement un fait d'observation très ancien (cf. l'article de J. F. PICHARD *Expérimentation et simulation probabiliste*). Jacques BERNOULLI, dans *Ars Conjectandi* (1713), lui a donné un sens précis : La probabilité que F_i s'écarte de p_i de plus qu'un ε donné tend vers 0 quand la taille de l'échantillon tend vers l'infini.

d'ailleurs être précisées. En effet, le nombre $N_i = nF_i$ d'éléments de l'échantillon qui sont de modalité M_i suit une loi binomiale $B(n, p_i)$, d'espérance np_i et de variance $np_i(1 - p_i)$. L'espérance mathématique de $F_i = \frac{N_i}{n}$ est donc p_i et son écart-type est $\sqrt{\frac{p_i(1 - p_i)}{n}}$. Le théorème de Bernoulli (énoncé dans la note 5 comme cas particulier de la loi faible des grands nombres) formalise ce phénomène de stabilisation[6].

Les probabilités p_i données dans le sous-modèle que l'on souhaite tester, peuvent provenir de différentes considérations. Dans cet article, nous nous limitons au cas où elles sont données a priori en fonction d'indications particulières (par exemple une hypothèse d'équiprobabilité entre les modalités M_i), seulement liées entre elles par la condition $\sum_{i=1}^{k} p_i = 1$.

II - Principe des tests du Khi-deux

Par principe, un test statistique consiste à se donner une hypothèse[7] (dite nulle) notée H_0, et à regarder si l'échantillon prélevé réalise un événement qui, si l'hypothèse H_0 était vraie, serait de probabilité relativement petite[8]. Dans ce cas, on refuse de considérer que cet événement est dû aux fluctuations d'échantillonnage et on préfère conclure que H_0 n'est pas acceptable (on dit qu'on rejette cette hypothèse, au profit d'une hypothèse alternative H_1). Ce faisant, on prend un risque de rejeter H_0 alors que seules les fluctuations d'échantillonnage sont responsables de la réalisation de cet événement, cependant peu probable.

Une loi modèle est donc donnée par une famille finie de probabilités (p) où l'on suppose que p_i est la probabilité que le caractère d'un élément pris au hasard dans la population soit dans la modalité M_i. Les M_i peuvent être les différentes qualités possibles du caractère *C* ou les différentes classes entre lesquelles on a regroupé les valeurs possibles de *C*, quand ce caractère est quantitatif. Un test d'adéquation consiste alors à regarder si, une distance d dans l'espace $\mathbb{R}^k$ étant choisie, la distance d((f) , (p)) de la loi (p) à la distribution de fréquences (f) observée dans un échantillon de taille n, est inférieure ou supérieure à une certaine valeur critique d_c.

---

[6] En 1908, Émile BOREL a démontré que la suite des fréquences f_i converge *presque sûrement* vers la probabilité p_i, c'est-à-dire que toutes les suites f_i possibles convergent (au sens des suites numériques) vers p_i, sauf pour un ensemble de suites lui-même de probabilité nulle (loi forte des grands nombres). Autrement dit, on n'a aucune chance d'observer une suite de fréquences qui ne convergerait pas vers p_i.

[7] Le mot *hypothèse* est à prendre ici au sens de *conjecture* que lui donnent les sciences expérimentales, et non pas au sens mathématique traditionnel de *prémisse*.

[8] Cf. l'article précédent *Introduction aux tests d'hypothèses, exemples.*

Cela revient à définir la valeur critique d_c par la condition que, *sous l'hypothèse* H_0, la variable aléatoire $D = d((F), (p))$, fonction des variables F_i, ne devrait dépasser d_c qu'avec une probabilité inférieure ou égale à un *seuil de signification* α. On voit que le calcul de cette probabilité de contrôle nécessite la connaissance, au moins approximativement, de la loi de D sous cette hypothèse H_0. C'est le cas (asymptotiquement[9]) de la distance des tests du Khi-deux (d'autres tests utilisent d'autres distances, comme celui de Kolmogorov-Smirnov[10]).

Dans un test du Khi-deux, on teste l'hypothèse :

- H_0 : « les probabilités des modalités M_i sont valablement modélisées par la loi de probabilité (p) »

contre l'hypothèse

- H_1 : « l'écart entre la distribution de fréquences (f) observée et la loi (p) n'est pas dû au hasard du prélèvement de l'échantillon ».

Ainsi, sous l'hypothèse H_0, les p_i sont les valeurs moyennes autour desquelles les fréquences F_i fluctuent.

Comme distance d dans $\mathbb{R}^k$, il peut sembler naturel de considérer la distance euclidienne dont le carré est $\sum_{i=1}^{k} (f_i - p_i)^2$, ce serait effectivement une mesure de la dispersion des F_i autour de leurs moyennes p_i. Or, pour les tests du Khi-deux, il s'avère plus intéressant de considérer la quantité $d^2((f), (p)) = \sum_{i=1}^{k} \frac{(f_i - p_i)^2}{p_i}$. La fonction d peut jouer le rôle d'une distance[11] puisqu'elle est positive et d'autant plus petite que les f_i sont proches des p_i. Mais le fait de pondérer les termes de cette somme par les $\frac{1}{p_i}$ a pour effet de la *normer*, en ce sens que pour n assez grand, la loi de D^2 ne dépend pratiquement plus des p_i, mais seulement de n et de k. C'est d'ailleurs le cas de sa moyenne, car pour tout n on a :

$$E(D^2) = E\left(\sum_{i=1}^{k} \frac{(F_i - p_i)^2}{p_i}\right) = \sum_{i=1}^{k} \frac{Var(F_i)}{p_i} = \sum_{i=1}^{k} \frac{p_i (1 - p_i)}{np_i} = \frac{k - 1}{n}.$$

Une autre raison est que l'on connaît cette loi, du moins asymptotiquement. C'est l'objet du Théorème du Khi-deux, dû à Karl PEARSON (1900), conséquence d'un grand théorème probabiliste, le théorème-limite central.

---

[9] Quand la taille n de l'échantillon tend vers l'infini, la loi de D converge vers une loi de χ^2 (i.e. la suite des fonctions de répartition des v. a. D converge simplement vers la fonction de répartition de ce χ^2).

[10] Voir par exemple [KAUFFMANN] p. 262.

[11] En fait cette « distance » d du Khi-deux n'est qu'une pseudo-distance, n'étant pas symétrique.

Théorème du Khi-deux :

Si, dans un échantillon aléatoire de taille n, F_i désigne la fréquence de la modalité M_i, et si p_i est la probabilité qu'un élément pris au hasard dans la population soit de modalité M_i, la suite des variables aléatoires $D_n^2 = n \sum_{i=1}^{k} \frac{(f_i - p_i)^2}{p_i}$ *converge en loi vers la loi de χ^2 à k–1 degrés de liberté.*

Pratiquement, cela signifie que pour n *assez grand*, et pour tout $Q > 0$, les probabilités $P(D_n^2 > Q)$ sont assez proches des $P(\chi^2_{k-1} > Q)$. Une table du Khi-deux (ou la fonction KHIDEUX.INVERSE d'Excel) donne le quantile Q pour un seuil α fixé, vérifiant $P(\chi^2_{k-1} > Q) = \alpha$.

On admet en général que n *assez grand* veut dire : pour tout i de 1 à k, $np_i(1-p_i) \geq 5$ (on trouve aussi $n > 30$ et $np_i \geq 5$), conditions qui, si elles ne sont pas vérifiées, impliquent de regrouper des modalités pour pouvoir appliquer ce théorème.

La loi de χ^2_ν est bien connue des probabilistes, elle est présentée dans les manuels universitaires de statistique[12] . Sa densité fait intervenir la fonction eulérienne de première espèce Γ (pour ν entier, $\Gamma(\nu) = (\nu - 1)!$) ; elle est donnée pour $x > 0$ par :

$$f_{\chi^2_\nu}(x) = \frac{x^{\frac{\nu}{2}-1}}{2^{\frac{\nu}{2}}\,\Gamma\left(\frac{\nu}{2}\right)} e^{-\frac{x}{2}}$$

et on a $E(\chi^2_\nu) = \nu$ et $Var(\chi^2_\nu) = 2\nu$. D'ailleurs, d'après le calcul ci-dessus, pour tout n, on a : $E(D_n^2) = k - 1$.

III - Pratique d'un test du Khi-deux

Dans la pratique, pour faire un test du Khi-deux, une loi modèle (p) étant donnée, il faut :

1) Vérifier que pour tous les i de 1 à k, on a $np_i(1-p_i) \geq 5$ (éventuellement regrouper des modalités).

2) Un seuil α étant donné, trouver dans la table des quantiles du χ^2_{k-1} à $k-1$ degrés de liberté la valeur χ^2_c telle que $P(\chi^2 > \chi^2_c) = \alpha$.

3) Expliciter la distribution (f) des fréquences présentées par l'échantillon observé.

[12] Voir par exemple [KAUFFMANN] p. 258 ou [DROESBEKE] p. 254 ou [VEYSSEYRE] p. 171.

4) Calculer *la distance* du Khi-deux : $d^2 = n \sum_{i=1}^{k} \frac{(f_i - p_i)^2}{p_i}$.

5) Conclure :
- si $d^2 \leq \chi_c^2$, on accepte l'hypothèse d'adéquation de la distribution de fréquences (f) des modalités M_i observée dans l'échantillon à la loi (p);
- si $d^2 > \chi_c^2$, on rejette cette adéquation avec un risque inférieur à α de se tromper[13].

Remarques :

i - Avec les effectifs $n_i = n f_i$ des éléments de l'échantillon de modalité M_i, l'expression de la distance du Khi-deux prend la forme traditionnelle :

$$d^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i},$$

ii - On prend souvent pour seuil $\alpha = 0{,}05$. Les quantiles χ_c^2 de la loi de χ_{k-1}^2, tels que $P(\chi_{k-1}^2 > \chi_c^2) = 0{,}05$ sont donnés par le tableau suivant :

nombre de modalités : k	2	3	4	5	6	7	8	9	10
degrés de liberté : k – 1	1	2	3	4	5	6	7	8	9
quantile χ_c^2 pour $\alpha = 0{,}05$	3,84	5,99	7,82	9,49	11,1	12,6	14,1	15,5	16,9

Pour les valeurs de k supérieures à 10, on peut utiliser la formule d'approximation[14] de Wilson-Hilferty :

$$\chi_c^2 \approx (k-1)\left(1 - \frac{2}{9(k-1)} + 1{,}645\sqrt{\frac{2}{9(k-1)}}\right)^3$$

Quand n est assez grand, on peut approcher la loi de D_n^2 par celle de χ_{k-1}^2 et conclure que pour 95 % des échantillons en moyenne, on a $\sum_{i=1}^{k} \frac{(f_i - p_i)^2}{p_i} < \frac{\chi_c^2}{n}$.

Cette inégalité montre que l'acceptation de l'adéquation de la distribution (f) à la loi (p) suppose une distance D_n^2 du Khi-deux qui tende vers 0 quand n tend vers l'infini, et indique sa liaison avec le nombre de modalités M_i via le quantile χ_c^2.

---

[13] i.e. Si l'adéquation est effectivement justifiée, la probabilité que l'échantillon observé conduise à la rejeter est majorée par α. Cela ne veut absolument pas dire que dans le cas où le modèle (p) ne serait pas adéquat, la probabilité de conclure au rejet de l'adéquation soit supérieure à $1 - \alpha$, elle vaut en fait $1 - \beta$, la puissance du test. (Cf. l'article précédent *Introduction aux tests d'hypothèses*.)

[14] Dans cette formule, le nombre 1,645 est le quantile correspondant à la probabilité 0,05 de la loi normale. Il existe d'autres formules d'approximation. Voir par exemple [VEYSSEYRE] p. 173.

iii - Dans le cas où $k = 2$ (épreuves de Bernoulli), on retrouve la condition de confiance pour l'estimation d'une probabilité p par la fréquence f des succès observés dans l'échantillon de taille n.

En effet, $\sum_{i=1}^{2} \frac{(f_i - p_i)^2}{p_i} = \frac{(f-p)^2}{p} + \frac{\left[(1-f)-(1-p)\right]^2}{1-p} = \frac{(f-p)^2}{p(1-p)}$ et la condition d'acceptation de H_0 au seuil $\alpha = 0{,}05$ du test du Khi-deux, donnée par $n\sum_{i=1}^{2} \frac{(f_i - p_i)^2}{p_i} < 3{,}84$, équivaut à la condition de confiance de niveau 0,95 : $|f - p| < 1{,}96\sqrt{\frac{p(1-p)}{n}}$.

IV - Adéquation à une loi équirépartie.

On veut savoir si on peut représenter la distribution de fréquences (f) entre les k modalités M_i par un modèle d'équiprobabilité : dans l'hypothèse H_0, les p_i sont donc supposées égales à $\frac{1}{k}$.

L'expression de *la distance* du Khi-deux devient :

$$d^2 = nk\sum_{i=1}^{k}\left(f_i - \frac{1}{k}\right)^2.$$

Une remarque peut alors simplifier la pratique de ce test et donner lieu un énoncé simple, accessible en terminale (cf. la fin de cet article).

Avec $\alpha = 0{,}05$, valeur courante, on peut voir dans la table qui précède ainsi que dans la formule de Wilson-Hilferty, que pour tout k, la valeur critique χ^2_c est comprise entre k et 2k. Cette simplification, relativement grossière, a l'avantage d'éviter d'utiliser la table du Khi-deux. Le test consiste alors à comparer la *distance* d^2 à 2k (en acceptant une perte de puissance), ce qui en simplifiant par k donne la décision pratique :

Si $n\sum_{i=1}^{k}\left(f_i - \frac{1}{k}\right)^2 > 2$, alors $d^2 > \chi^2_c$ et on rejette l'hypothèse d'équirépartition du caractère entre ses k modalités, avec un risque inférieur à 0,05.

Avec les effectifs $n_i = n f_i$, la condition de rejet de l'équirépartition s'écrit aussi :

$$\sum_{i=1}^{k}\left(n_i - \frac{n}{k}\right)^2 > 2n,$$

Exemple

On veut tester si un dé est régulier ; autrement dit, si on peut accepter que la probabilité qu'il tombe sur l'une ou l'autre face est égale à $\frac{1}{6}$. On décide de le lancer 600 fois. On observe les effectifs n_1, n_2, …, n_6 d'apparition des six faces. On calcule $\sum_{i=1}^{6} (n_i - 100)^2$. Si ce nombre dépasse 1 200, on rejette l'hypothèse que le dé est régulier.

V - Les enjeux didactiques

1 - Enjeu de formation de l'équiprobabilité

L'enjeu de formation est capital, puisqu'il concerne la démarche de *modélisation*. Le programme actuel de terminale S se limite aux situations d'équiprobabilité, ce qui paraît judicieux. D'une part, c'est plus simple : la quantité $\sum_{i=1}^{k} \left(f_i - \frac{1}{k}\right)^2$ est plus naturelle que la quantité $\sum_{i=1}^{k} \frac{(f_i - p_i)^2}{p_i}$ et la majoration de χ^2_c par 2k évite le recours aux tables. D'autre part, c'est le cas où l'approximation de la loi de $n \sum_{i=1}^{k} \frac{(f_i - p_i)^2}{p_i}$ par celle de χ^2_{k-1} est la meilleure[15].

Surtout, il est essentiel de faire comprendre que l'équiprobabilité est une hypothèse forte, qui ne va pas de soi. On peut s'étonner à ce propos que dans les manuels, beaucoup d'énoncés supposent une loi uniforme sans le dire : on parle d'un dé sans préciser qu'il est régulier, d'une pièce sans dire qu'elle est équilibrée[16]… c'est comme si en géométrie on proposait un exercice sur un triangle équilatéral en omettant de dire qu'il est équilatéral ! La remarque vaut bien sûr pour les jeux de cartes (bien battus), les tirages (au hasard), etc.

Dans l'exemple ci-dessus, modéliser un dé réel consiste à le représenter par un dé régulier[17]. Il faut souligner que le dé régulier n'existe pas dans la Nature, pas plus que le triangle équilatéral. Il s'agit d'idéalités mathématiques sur lesquelles on

---

[15] Voir plus loin.

[16] On peut remarquer d'ailleurs que l'hypothèse d'équiprobabilité est plus naturelle pour un dé, conçu en principe pour être régulier, que pour une pièce de monnaie qui n'est pas fabriquée pour cela et dont il existe de multiples variétés !

[17] Les issues 1, 2, …, 6 ayant été définies, un *modèle* du dé est donné par une loi de probabilités $(p_1, p_2, \ldots, p_6)$. Le programme se limite à l'équiprobabilité, mais bien entendu on pourrait envisager pour un dé un tout autre modèle, par exemple $(1/8, 1/8, 1/8, 1/8, 1/8, 3/8)$. Notons à ce propos que le modèle intègre non seulement les caractéristiques physiques du dé mais aussi la façon dont il est lancé.

peut raisonner. La question : « ce dé est-il régulier ? » signifie en fait : « le modèle du dé régulier est-il adapté à ce dé ? ».

Dès lors on conçoit que, selon la précision de la mesure, c'est-à-dire ici la taille de l'échantillon et le degré de confiance choisi, la réponse puisse être oui ou non. De même qu'à la question : « Le triangle Paris-Bordeaux-Grenoble est-il équilatéral ? », on peut répondre oui ou non selon qu'on mesure les distances à 100 km près ou à 10 km près.

On peut même pousser plus loin l'analogie avec la géométrie : la question ci-dessus n'a plus de sens si on mesure les distances au km près, car le modèle ponctuel pour chaque ville n'est alors plus valide. De la même façon, on peut considérer que pour un dé réel, il n'y a pas de sens de définir les probabilités d'apparition des six faces à 10^{-4} près, car le nombre de jets qui serait nécessaire pour estimer ces probabilités[18] risquerait d'altérer ses caractéristiques physiques, rendant impossible la répétition de l'épreuve dans les mêmes conditions.

Dernier argument, mais non le moindre, en faveur de l'équiprobabilité : toute épreuve d'univers fini peut se ramener à une situation d'équiprobabilité[19]. Par exemple, jeter un dé truqué où le 6 a trois fois plus de chances d'apparaître que chacune des 5 autres valeurs équivaut à tirer une boule dans une urne de Bernoulli contenant 8 boules marquées 1, 2, 3, 4, 5, 6, 6, 6. C'est d'ailleurs cette propriété qui permet les simulations faites en classe de seconde qui doivent amener à distinguer ce qui est équiprobable et ce qui ne l'est pas[20]. De ce point de vue, il paraîtrait très utile de faire tester les générateurs de *nombres au hasard* fournis par les tableurs ou les calculatrices, de façon à faire comprendre qu'ils simulent bien une loi uniforme. Faute de quoi on risque de gros malentendus lors des simulations (par exemple simuler la somme de deux dés réguliers par des nombres au hasard entre 2 et 12).

2 - Expérimentations en classe

Cela dit, la théorie présentée ci-dessus n'est pas accessible à un élève de terminale.

C'est pourquoi le document d'accompagnement du programme propose la démarche suivante :

---

[18] La théorie conduit à $n > 10^8$: peu de dés résisteraient à plus de 100 millions de jets (ce qui d'ailleurs, à raison d'un jet par seconde, durerait plus de trois ans) !

[19] Cela suppose les p_i rationnels, ce qui ne restreint pas vraiment le champ d'application de l'urne de Bernoulli à des situations concrètes : tout irrationnel pouvant être approché d'aussi près qu'on veut par un rationnel et, tout modèle étant une approximation de la réalité, on peut toujours envisager un modèle où les p_i sont rationnels. On trouvera dans [THIÉNARD] une réflexion approfondie sur l'utilisation des urnes pour l'enseignement des probabilités au lycée.

[20] Par exemple quand on jette deux dés réguliers indépendants, c'est parce que les couples sont équiprobables que les sommes ne le sont pas.

- Dans un premier temps : faire *simuler* au tableur une épreuve à k issues équiprobables, calculer $n\sum_{i=1}^{k}\left(f_i-\frac{1}{k}\right)^2$ pour un grand nombre d'échantillons de taille n donnée, écarter les 5 % d'échantillons qui fournissent les valeurs les plus élevées et conserver comme seuil critique la plus grande valeur des échantillons restants.
- Dans un deuxième temps : appliquer le test pour une situation d'équiprobabilité supposée.

Ainsi, pour tester la régularité d'un dé, on proposera comme ci-dessus de le lancer 600 fois, puis de calculer $600\sum_{i=1}^{k}\left(f_i-\frac{1}{6}\right)^2$, pour rejeter l'hypothèse de régularité si ce nombre est trop grand.

Mais pour savoir quel est le seuil critique qui permettra de trancher, on simulera au tableur 600 jets d'un dé régulier. On itérera par exemple 1 000 fois un tel échantillon de taille 600. On écartera les 50 échantillons qui fournissent pour $600\sum_{i=1}^{k}\left(f_i-\frac{1}{6}\right)^2$ les valeurs les plus élevées, et on prendra comme seuil critique la plus grande valeur des échantillons restants[21].

Cette démarche expérimentale, intéressante en soi, n'est pas si facile à mettre en œuvre :

i - Elle suppose une bonne maîtrise de la simulation sur tableur ou calculatrice. Cette compétence est en principe acquise en seconde, ce qui rend d'ailleurs cohérente la succession des programmes seconde-première-terminale en statistique et probabilités. Il y a cependant quelques difficultés, dues entre autres à la mauvaise compréhension de ce qu'est un générateur de *chiffres au hasard*[22]. Est-ce pour cela que le programme parle de « *résultats de simulation qu'on lui fournit* » ? Si ce n'est pas l'élève qui fait lui-même la simulation, l'intérêt s'en trouve réduit.

ii - Elle privilégie la quantité $n\sum_{i=1}^{k}\left(f_i-\frac{1}{k}\right)^2$, ce qui s'explique par le désir de faire apparaître un invariant, mais rend plus difficile la compréhension.

[21] En fait les commentaires préconisent de travailler au risque de 10 %, ce qui est peu courant en statistique et rend plus difficile la formulation du théorème énoncé à la fin de cet article.

[22] Cf. l'article de Bernard Parzysz *Quelques questions à propos des tables et des générateurs pseudo-aléatoires* dans ce même volume.

Bien sûr il est équivalent de dire que $n\sum_{i=1}^{k}\left(f_i - \frac{1}{k}\right)^2$ est inférieur à 2, ou que $\sum_{i=1}^{k}\left(f_i - \frac{1}{k}\right)^2$ est inférieur à $\frac{2}{n}$. Mais c'est la deuxième écriture qui fait sens, puisqu'elle indique que la précision de l'approximation augmente avec n. Prendre conscience de cela suppose d'ailleurs de faire varier la taille des échantillons, ce qui alourdit beaucoup ce travail expérimental.

iii - Elle superpose deux niveaux d'échantillonnage : on répète l'épreuve pour constituer un échantillon de taille n, puis on répète cet échantillonnage pour comparer de nombreux échantillons entre eux. Cet empilement constitue un obstacle important à la compréhension de ce que l'on fait, alors que le deuxième niveau d'échantillonnage n'est là que pour obtenir expérimentalement des résultats de théorèmes probabilistes.

Surtout, cette approche soulève des questions difficiles :

iv - Elle suppose qu'on observe un nombre suffisant d'échantillons (combien ?) pour que la proportion de 95 % ait un sens statistique.

v - Elle suppose que le générateur de nombres aléatoires du tableur est parfait. Or précisément d'après la théorie ci-dessus, il ne peut pas l'être. Tout au plus peut-on le tester, notamment à l'aide du Khi-deux, à un niveau de confiance donné. Il y a là comme un cercle vicieux.

vi - Elle risque de faire croire que le seuil critique dépend de l'expérimentateur. Il est déjà difficile de faire comprendre qu'un même dé puisse être jugé régulier au vu d'un échantillon, non régulier au vu d'un autre échantillon. Il ne faudrait pas obscurcir cette compréhension en faisant croire que le seuil critique n'est pas le même en terminale S et en terminale ES.

3 - Démarche didactique

C'est pourquoi il semble qu'il faudrait ramener cette démarche expérimentale à ce qu'elle doit être en mathématiques : une activité d'introduction, indispensable pour motiver les notions à introduire, mais qui nécessite une institutionnalisation, c'est-à-dire l'énoncé d'un théorème. Que dirait-on d'un professeur de mathématiques qui, ayant fait mesurer les côtés de divers triangles rectangles et constater que la somme des carrés des côtés est à peu près égale au carré de l'hypoténuse, ne jugerait pas utile d'énoncer le théorème de Pythagore ?

En énonçant un théorème (admis bien entendu), on dépasse le questionnement ci-dessus. L'activité ne prétend rien démontrer, elle a un rôle heuristique, fondamental mais limité : faire manipuler et comprendre les résultats théoriques qui viennent ensuite.

Mais quel théorème énoncer ?

Pour ne pas trop compliquer une notion déjà délicate, on pourrait se limiter à une seule valeur pour le seuil de signification, par exemple 0,05. Cela permettrait d'appeler *rare* ou *exceptionnel*, par définition, un événement qui a une probabilité inférieure à 0,05.

Mais on bute ici sur une difficulté, déjà rencontrée en seconde avec la fourchette de sondage. Énoncer un théorème suppose de pouvoir donner un résultat rigoureux et démontré, donnant la valeur de n à partir de laquelle la probabilité dépasse toujours 0,95.

Or à notre connaissance les études accessibles ne fournissent pas actuellement un tel théorème. Elles indiquent seulement que la vitesse de convergence est meilleure quand les p_i ne sont pas trop proches de 0 ou de 1. On admet souvent comme critères[23] $n > 30$, $np_i \geq 5$ pour tout i. Dans le cas de l'équiprobabilité, ces conditions se ramènent à $n > 30$ et $n \geq 5k$. Le commentaire des programmes stipule $n \geq 100$. Mais il s'agit simplement d'habitudes des statisticiens : aucun livre ne donne la validité de l'approximation quand ces conditions sont remplies.

En fait il faudrait étudier la loi (discrète) de $n\sum_{i=1}^{k}\left(f_i - \frac{1}{k}\right)^2$ pour pouvoir énoncer un véritable théorème[24], garantissant un niveau de confiance de 0,95.

Notons que cette absence de précision justifie, dans le cas de l'équiprobabilité, de prendre $\frac{2}{n}$ comme seuil critique pour $\sum_{i=1}^{k}\left(f_i - \frac{1}{k}\right)^2$: raffiner en utilisant les tables du Khi-deux constituerait un gain de précision illusoire, puisque la variable $k\,n\sum_{i=1}^{k}\left(f_i - \frac{1}{k}\right)^2$ ne suit qu'approximativement une loi de χ^2.

En attendant que les chercheurs nous fournissent un théorème plus précis, proposons l'énoncé vulgarisé[25] suivant, accessible en terminale.

Soit une épreuve ayant k issues équiprobables. On la répète n fois et on note f_1, f_2, , f_k les fréquences d'apparition des différentes issues. Si n est assez grand, il est exceptionnel que $\sum_{i=1}^{k}\left(f_i - \frac{1}{k}\right)^2$ dépasse $\frac{2}{n}$.

---

[23] Voir par exemple [DROESBEKE], p. 354.

[24] Dans le cas $k = 2$, une justification numérique a été publiée tout récemment par Louis-Marie BONNEVAL (voir le Bulletin APMEP n° 436, pages 732-733). La démonstration complète de la validité de ces conditions a été obtenue par Christian MAILLARD, professeur à Avignon.

[25] Cet énoncé simplifié d'un théorème de K. Pearson est proposé par Louis-Marie BONNEVAL.

Volontairement les expressions *n est assez grand* et *il est exceptionnel* n'ont pas été précisées. On pourra signaler en commentaire que si $n > 30$ et $n \geq 5k$, il y a environ un échantillon sur 20 pour lequel $\sum_{i=1}^{k} \left(f_i - \frac{1}{k} \right)^2$ dépasse $\frac{2}{n}$.

Armés de cet énoncé, les élèves pourront dès lors adopter un point de vue critique sur l'hypothèse d'équiprobabilité, sans devoir, chaque fois qu'une situation se présente, commencer par en simuler de multiples échantillons. Notamment, ils pourront contrôler les générateurs de nombres au hasard qui sont l'outil des simulations. Ils pourront avoir une véritable activité mathématique, c'est-à-dire raisonner. Surtout, et c'est là l'objectif principal, on peut espérer qu'ils auront une première idée de ce que peut être un test statistique.

Bibliographie :

CARNEC H., SEROUX R., DAGOURY J.M., THOMAS M., (2000) : *Itinéraires en statistiques et probabilités*, Ellipses.

DROESBEKE J.-J., (1997) : *Eléments de statistique*, Ellipses, 3ème édition.

ENGEL A., (1990) : *Les certitudes du hasard*, Aléas.

KAUFFMANN, P., (1994) : *Statistique, Information, Estimation, Tests*, Dunod.

SAPORTA G., (1990) : *Probabilités, analyse des données et statistiques*, Technip.

SPIEGEL M. R., (1993) : *Théorie et applications de la statistique*, Mc-Graw-Hill, 2éme édition.

THIÉNARD J.-C., (1993) : *À propos de l'enseignement du calcul des probabilités*, IREM de Poitiers.

VEYSSEYRE R., (2000) : *Statistique et probabilités pour l'ingénieur*, Dunod.

Annexes


4
3
2

Publications des IREM et de l'APMEP sur l'enseignement de la statistique et des probabilités

Publications de la Commission Inter-IREM *Statistique et Probabilités*

Éditées par l'IREM de Rouen :

– *Actes de l'université d'été de statistiques inférentielles*, La Rochelle : 1er au 5 septembre 1992, 238 p.

– *Actes de l'université d'été de statistiques inférentielles*, Rouen : 29 août au 2 septembre 1994, 252 p.

– Actes de l'université d'été de Probabilités, Metz 26 au 31 août 1996, 322 p.

Éditées par l'IUT de Metz :

– *Modélisation en probabilités*, octobre 1997, 60 p.

– *Aide pour l'enseignement des statistiques*, octobre 1997, 50 p.

Éditée par l'IREM de Reims :

– *Enseigner les probabilités au lycée*, octobre 1997, 464 p.

Éditée par les Presses Universitaires Franc-Comtoises :

– *Autour de la modélisation en probabilités,* avril 2001, 260 p.

Éditée par l'APMEP :

– *Probabilités au lycée*, février 2003, brochure n° 143, 184 p.

Publications de la Commission Inter-IREM *Histoire et Epistémologie des Mathématiques*

Mathématiques au fil des âges, éd. Gauthiers-Villars 1987,

– *Calcul des probabilités*, collectif, chapitre 5, pp. 211-234.

Actes du 7ème colloque, Besançon, 1989 : *La Démonstration Mathématique dans l'Histoire*, éd. IREM de Besançon, 1990 :

– *Argumentation et démonstration : à quoi sert la démonstration de "la loi des grands nombres" de Jacques Bernoulli*, Norbert MEUSNIER, pp. 81-97.

Actes du 9ème colloque, Brest, 1992, *Histoire d'infini*, éd. IREM de Brest, 1994 :

– *Huygens : l'espérance et l'infini*, Denis LANIER, pp. 555-577.

Actes du 10ème colloque, Caen, 1994, *La Mémoire des Nombres*, éd. IREM de Caen, 1997 :

– *Buffon et le problème de l'aiguille*, Frédéric MÉTIN, pp. 343-360.

Actes de la 6ème Université d'été interdisciplinaire sur l'histoire des mathématiques, Besançon, 1995, thème II : *Histoire des probabilités et des statistiques*, éd. IREM de Besançon, 1996 :

– *De Cassini à Gauss : du calcul d'erreurs aux probabilités*, Anne BOYÉ et Xavier LEFORT, pp. 239-258.

– *La Loi des grands nombres*, Denis LANIER et Didier TROTOUX, pp. 259-294.

– *Quelques anciens problèmes de probabilités*, Michèle Lacombe et Henry PLANE, pp. 295-304.

– *Formes et significations des probabilités chez Cournot : la fortuité des décimales de π*, Thierry MARTIN, pp. 305-318.

Actes du 11ème colloque, Reims, 1996, *Analyse & démarche analytique,* éd. IREM de Reims, 1998 :

– *Le jeu du Treize, un essai d'analyse d'un jeu de hasard*, Patrick PERRIN, pp. 205-229.

– *La formule de Stirling*, Denis LANIER et Didier TROTOUX, pp. 231-286.

– *Abraham de Moivre*, Gilbert MAHEUT, pp. 373-386.

Actes de la 7ème Université d'été interdisciplinaire sur l'histoire des mathématiques, Nantes, 1997, *Contribution à une approche historique de l'enseignement des mathématiques,* éd. IREM des Pays de Loire, 1999 :

– *On the historical phenomenology of probabilistic concepts - from the didactical point of view*, Ewa LAKOMA, p. 439.

Actes de la 8ème Université d'été interdisciplinaire sur l'histoire des mathématiques, Louvain, 1999, *Histoire et épistémologie dans l'éducation mathématique « de la maternelle à l'université »,* éd. Université Catholique de Louvain, Belgique, 2001.

– *Histoire des mathématiques du chaos et épistémologie du hasard*, Jacqueline GUICHARD, vol 1, p. 213.

– *On the duality of probability concept - from the epistemological point of view*, Ewa LAKOMA, vol. 2, p. 395.

Actes du 14ème colloque, Orléans, 2002, *Histoire de probabilités et de statistiques*, coord. Evelyne BARBIN, Jean-Pierre LAMARCHE, Ellipse, 2004 :

– *Le problème des partis avant Pacioli*, Norbert MEUSNIER, pp. 3-24.

– *Huygens et ses lecteurs : le 5ème exercice*, Denis LANIER et Didier TROTOUX, pp. 25-54.

– *La portée physique et sociale de la règle de Bayes*, Jean-Pierre CLÉRO, pp. 55-73.

– *Le joueur et le banquier. Sur une correspondance des frères Huygens*, Bernard PARZYSZ, pp. 77-90.

– *Tables de natalité, tables de mortalité « À tables ! »*, Henri Plane, Frédéric MÉTIN, Patrick GUYOT, pp. 91-118.
– *La démonstration par Jacques Bernoulli de son théorème*, Michel HENRY, pp. 121-140.
– *La théorie des erreurs (1750-1820), enjeux, problématiques, résultats*, Michel ARMATTE, pp. 141-160.
– *Statistique et modèles probabilistes de Fisher à Havelmoo*, Martin ZERNER, pp. 161-172.
– *La controverse antique sur les futurs contingents*, Michèle VILLETARD TAINMONT, Joëlle DELATTRE, pp. 175-196.
– *Laplace et la Théorie analytique des probabilités : itinéraires de découverte*, Jean-Pierre LUBET, pp. 197-224.
– *Cournot. Statistique et raison des choses*, Thierry MARTIN, pp. 225-234.
– *Sur l'histoire de l'enseignement des probabilités et des statistiques*, Norbert MEUSNIER, pp. 237-274.
– *Galilée ou Descartes ? Etude d'un scénario d'introduction historique au calcul des probabilités*, Eric BUTZ, pp. 275-296.

Publications de la Commission Inter-IREM *Lycées Techniques*, éditées par l'IREM de Paris-Nord

– *A propos* de *fiabilité*, Brochure n° 48.
– *Les* plans *d'expérience pour le BTS chimiste*, J.-L. PIEDNOIR, J.-P. BENICHOU, brochure n° 88.
– *Simulation* d'expériences *aléatoires. une expérience du hasard de la première au BTS sur calculatrice et ordinateur,* P. DUTARTE, C. KERN, M.-F. NOUGUÈS, G. SAINT-PIERRE, B. VERLANT, brochure n° 93, 1998.
– *Simulation et statistique en Seconde,* P. Dutarte, C. KERN, D. ARBRE, I. BRUN, F. DELZONGLE, C. DHERS, A. LADUREAU, J.-L. LANGON, M.-F. NOUGUÈS, G. SAINT-PIERRE, B. VERLANT, brochure n° 102, 2000.
– *Enseigner la* statistique *au lycée : des enjeux aux méthodes,* J.-L. PIEDNOIR, P. DUTARTE, brochure n° 112, 2001.
– *La statistique inférentielle en quatre séances,* P. DUTARTE, C. KERN, D. ARBRE, I. BRUN, F. DELZONGLE, C. DHERS, L. MAZO, M.-F. NOUGUÈS, G. SAINT-PIERRE, B. VERLANT, brochure n° 118, 2002.
– *Le nouveau programme de statistique et probabilités au lycée,* P. DUTARTE, D. ARBRE, I. BRUN, F. DELZONGLE, C. DHERS, C. KERN, L. MAZO, M.-F. NOUGUÈS, B. VERLANT, brochure n° 124, 2003.

Articles sur l'enseignement des probabilités et de la statistique parus dans la revue Repères-IREM

- *L'enseignement des probabilités dans le programme de première*, Annie HENRY et Michel HENRY, n° **6**, janvier 1992.
- *Des statistiques aux probabilités : exploitons les arbres,* Bernard PARZYSZ, n° **10**, janvier 1993.
- *Paradoxes et lois de probabilités,* Michel HENRY et Henri LOMBARDI, n° **13**, octobre 1993.
- *L'enseignement du calcul des probabilités dans le second degré : perspectives historiques, épistémologiques et didactiques,* Michel HENRY, n° **14**, janvier 1994.
- *L'introduction du concept de probabilité conditionnelle : avantages et inconvénients de l'arborescence,* André TOTOHASINA, n° **15**, avril 1994.
- *L'apprenti fréquentiste,* Jean-Claude DUPERRET, n° **21**, octobre 1995.
- *Arbres et tableaux de probabilité : analyse en termes de registres de représentation,* Suzette ROUSSET-BERT, Claire DUPUIS, n° **22**, janvier 1996.
- *Pourquoi il ne faut pas laisser de côté les chapitres de statistiques au collège,* Jean Claude GIRARD, n° **23**, avril 1996.
- *Approche épistémologique et diverses conceptions de la probabilité,* Jean-François PICHARD, n° **32**, juillet 1998.
- *Expérimenter et simuler en classe,* Michèle GANDIT, Claire HELMSTETTER, n° **32**, juillet 1998.
- *Attention ! Un modèle peut en cacher un autre,* Hubert RAYMONDAUD, Michel HENRY, n° **32**, juillet 1998.
- *Chronique d'une correspondance probablement apocryphe,* Jacques VERDIER, Pol LE GAL, André VIRICEL, Bernard PARZYSZ, Gilberte PASCAL, n° **32**, juillet 98.
- *Qu'est-ce que le hasard ? comment le mathématiser ?* Claude CHRÉTIEN, Dominique GAUD, n° **32**, juillet 1998.
- *Un problème de dés en Terminale*, Martine BÜHLER, n° **32**, juillet 1998.
- *A bas la moyenne !*, Jean Claude GIRARD, n° **33**, octobre 1998.
- *La moyenne : un concept inexploité, d'une richesse exceptionnelle,* Linda GATTUSO, n° **34**, janvier 1999.
- *Heurs et malheurs du su et du perçu en statistique. Des données à leurs représentations graphiques,* Bernard PARZYSZ, n° **35**, avril 1999.
- *Le professeur de mathématiques doit-il enseigner le modélisation ?,* Jean Claude GIRARD, n° **36**, juillet 1999.

- *L'introduction des probabilités au lycée : un processus de modélisation comparable à celui de la géométrie,* Michel HENRY, n° **36**, juillet 1999.
- *La fonction de répartition. Pour quoi faire ?* Pascale POMBOURCQ, n° **38**, janvier 2000.
- *Probabilités, suites numériques et programmation,* Michel BOURGUET, n° **41**, octobre 2000.
- *Quelle place pour l'aléatoire au collège ?*, Jean Claude GIRARD, Michel HENRY, Bernard PARZYSZ, Jean-François PICHARD, n° **42**, janvier 2001.
- *Une enquête statistique au service de la proportionnalité,* Daniel GROS, n° **44**, juillet 2001.
- *Vache folle, probabilités, réalités et… pesanteur(s),* Gérard KUNTZ, n° **45**, octobre 2001.
- *Simulation et modélisation : étude d'un exemple,* Joëlle FONTANA et Maryse NOGUÈS, n° **46**, janvier 2002.
- *La simulation en statistique,* Philippe DUTARTE, n° **47**, avril 2002.
- *A propos de camemberts, ou l'art de manipuler l'information,* François GOULETQUER, n° **48**, juillet 2002.
- *Une approche de la normalité à l'aide de la planche de Galton,* Anne CROUZIER, n° **48**, juillet 2002.
- *Moyenne, médiane, écart-type, quelques regards sur l'histoire pour éclairer l'enseignement des statistiques au lycée,* Anne BOYÉ et Marie-Céline COMAIRAS, n° **48**, juillet 2002.
- *Sur l'enseignement de la statistique en Communauté française de Belgique,* Jacques BAIR et Gentiane HASBROECK, n° **48,** juillet 2002.
- *Des lois continues, pourquoi et pour quoi faire ?,* Michel HENRY, n° **51**, avril 2003.
- *Les statistiques ; de leur utilisation dans le domaine public en remontant vers leurs sources,* René MULET-MARQUIS, n° **55**, avril 2004.
- *La liaison statistique-probabilités dans l'enseignement,* Jean Claude GIRARD, n° **57,** octobre 2004.

Articles sur l'enseignement des probabilités et de la statistique parus dans le bulletin vert de l'APMEP de 1998 à 2005

- *Simulation d'un exercice de probabilité,* Bruno LEVAT, Daniel VAGOST, n° 418, septembre-octobre 1998.
- *« Les maths, c'est pas la réalité ! » ou De la modélisation en mathématiques,* Jean Claude GIRARD, Bernard PARZYSZ, n° 418, septembre-octobre 1998.
- *Un enseignement des probabilités en premier cycle de la filière AES*, Yves DUCEL, Hombeline LANGUEREAU, n° 420, janvier-février 1999.
- *Des approches variées pour un même phénomène : la datation au radiocarbonne,* Bernard PARZYSZ, n° 421, mars-avril 1999.
- *À propos de l'enseignement de la statistique au lycée,* Claudine ROBERT, dossier Statistique et probabilités, n° 425, novembre-décembre 1999.
- *Enseigner les statistiques en seconde,* Rémy COSTE, dossier Statistique et probabilités, n° 425, novembre-décembre 1999.
- *La géométrie au service de la statistique,* Jean-Louis PIEDNOIR, dossier Statistique et probabilités, n° 425, novembre-décembre 1999.
- *Les probabilités du bac S de 1998,* Jean-Pierre GRANGÉ, dossier Statistique et probabilités, n° 425, novembre-décembre 1999.
- *Nos élèves refont l'histoire des probabilités,* Nicole VOGEL, dossier Statistique et probabilités, n° 425, novembre-décembre 1999.
- *Un choix d'ouvrage pour enseigner le programme de statistique de seconde*, Bernard PARZYSZ, n° 427, mars-avril 2000.
- *Intervalles de confiance ?*, Louis-Marie BONNEVAL, n° 427, mars-avril 2000.
- *Que peut-on faire avec le programme de statistique de collège*, Bernard PARZYSZ, n° 428, avril 2000
- *Deux ou trois petites choses que je sais de la médiane*, Jacques VERDIER, n° 430, septembre-octobre 2000.
- *Entre réel et virtuel, la simulation en statistique*, Bernard EGGER, n° 434, mai-juin 2001.
- *Deux ou trois choses que je sais des quartiles et des boîtes à moustaches*, Jacques VERDIER, n° 435, septembre-octobre 2001.
- *De l'espace aux statistiques (et retour)*, Thierry LAPOUGE, n° 439, mars-avril 2002.
- *Sur la pulsation probabiliste*, Laurent MAZLIAK, n° 440, mai-juin 2002.

- *Mathématiques et tableur au lycée : le problème du duc de Toscane*, Virginie MAITROT, n° 440, mai-juin 2002.
- *Test d'équirépartition : qui a dit khi-deux ?*, Louis-Marie BONNEVAL, n° 441, septembre-octobre 2002.
- *Maximum d'un écart-type d'une série statistique bornée et docimologique*, Serge CHESNEY, n° 443, novembre-décembre 2002.
- *Petit essai sur les tirages dans une urne*, Yves-Noël HAUBRY, n° 443, novembre-décembre 2002.
- *Simulation d'un sondage. Fourchettes d'échantillonnage et intervalles de confiance*, Michel HENRY, n° 444, janvier-février 2003.
- *Quelques calculs de probabilités dans la vie courante*, Robert FERRÉOL, n° 447, septembre-octobre 2003.
- *Politique nataliste : Analyse d'un devoir sur une "régulation des naissances"*, Jacques VERDIER, n° 447, septembre-octobre 2003.
- *Peut-on imiter le hasard*, Nicole VOGEL, n° 451, mars-avril 2004.
- *Lois continues en Terminale S, quelle approche ?*, Michel HENRY, n° 454, octobre 2004.
- *Test d'équirépartition, quel risque d'erreur ?*, Louis-Marie BONNEVAL, n° 456, janvier-février 2005.
- *Géométrie et probabilités : une même démarche de modélisation ?*, Louis-Marie BONNEVAL, n° 456, janvier-février 2005.
- *Coïncidences des dates anniversaires*, Jean-François KENTZEL, n° 457, mars-avril 2005.

Brochures de l'APMEP

- *Petite histoire du calcul des probabilités,* Bernard Bru, Fragments d'histoire des mathématiques, tome 1, brochure n° 41, 1981.
- *En passant par le hasard*, Gilles PAGÈS, Claude BOUZITAT, éd. Vuibert, codiffusion APMEP, brochure n° 907, 2000.
- *Les statistiques en classe de seconde*, Pascale POMBOURCQ, n° 138, 2001.
- *Enseigner la statistique au lycée : des enjeux aux méthodes*, Jean-Louis PIEDNOIR, Philippe DUTARTE, éd. IREM Paris-Nord, codiffusion APMEP, brochure n° 820, 2001.
- *Contes et décomptes de la statistique,* Claudine ROBERT, éd. Vuibert, codiffusion APMEP, brochure n° 930, 2003.
- *Probabilités au lycée*, 2ème édition, CII Statistique et Probabilités, n° 143, 2003.

Brochures des IREM publiées depuis 1998

La liste des brochures publiées de 1991 à 1996 figure en annexe 3 dans le livre *Enseigner les probabilités au lycée*, édité en juin 1997 par l'IREM de Reims.

IREM d'Aquitaine

- *L'esprit des lois continues ou quelques aspects du calcul des probabilités au lycée,* E. BARBAZO, J.-M. BOUSCASSE, R. POMÈS, J. PUYOU, M. PUYOU, P.-H. TERRACHER, 2003.

IREM de Clermont-Ferrand

- *Enseignement des probabilités-statistiques en lycée (stage IUFM),* D. ARBRE, A. CORPART, R. NOIRFALISE, P. PHILIPPE, 2000.
- *Une application industrielle des statistiques : la carte de contrôle,* D. ARBRE, A. CORPART, G. FLEURY, N. LASSALLE, 2001.
- *Une application industrielle des statistiques : la loi exponentielle,* D. ARBRE, J.-L. CHAMPOMMIER, A. CORPART, G. FLEURY, N. LASSALLE, 2003.

IREM de Dijon

- *Statistiques en seconde et… un peu après*, M. PLATHEY, F. GOUTLETQUER, M. BRIDENNE, D. GARDES, F. MARCHIVIE, A. JEBRANE, 2000.

IREM de Franche-Comté

- *Arbres et probabilités*, J.-P. GRANGÉ, 1999.

IREM de Lille

- *Probabilités et statistique : autour de la loi exponentielle,* R. MOCHÉ, 2000.
- *Itinéraire d'un calcul annoncé. Quelques activités de la seconde au premier cycle universitaire,* M. GOUY, G. HUVENT, A. LADUREAU, 2003.
- *Le Hasard,* actes des journées académiques des 15 et 16 avril 2004, R. MOCHÉ (coord.), conférences :
    1. *Application des probabilités dans l'industrie : calculs de fiabilité des systèmes*, Marc BOUISSOU (EDF et CNRS)
    2. *L'homme multidimensionnel et l'opinion publique, quelques réflexions sur les statistiques et les sciences sociales*, Nicolas BOULEAU (Ecole des Ponts et Chaussées, Paris)
    3. *Exposé général sur la statistique : peut-on modéliser le hasard ?*, Michel CARBON (ENSAI, Rennes)
    4. *La notion de probabilité: évolution historique et applications contemporaines*, Michel HENRY (IREM de Besançon)
    5. *Modélisation en probabilités : un modèle peut-il en cacher un autre ?,* Pierre-Henri TERRACHER (IREM de Bordeaux)

IREM de Lorraine

– *L'enseignement des probabilités au collège et au lycée : exemples européens et propositions*, F. CHAIBAI, B. PARZYSZ, D. VAGOST, J. VERDIER, 2001.

IREM de Lyon

– *Enseigner la statistique du CM à la seconde. Pourquoi ? Comment ?*, J. C. GIRARD, D. GROS, P. PLANCHETTE, J.-C. RÉGNIER, R. THOMAS, 1998.

IREM de Montpellier

– *Des statistiques à la pensée statistique,* N. BASCOU, A. BERNARD, M.-C. COMBES, J.-C. DUPERRET, J. FONTANA, A. GANNOUN, M. HENRY, M. JANVIER, M.-F. JOZEAU, M. LACAGE, M. NOGUÈS, J.-M. RAVIER, M. ROCHE, N. SABY, J. SALLES, M. SAUTER, M. SECO, L. TROUCHE, C. VERGNE, 2001.

IREM d'Orléans-Tours

– *Réflexions sur l'enseignement des statistiques au Lycée,* J.-P. GERBAL, J.-P. LAMARCHE, J. LUCY, M. ROSÉ, H. VASSEUR, Cahiers de l'IREM d'Orléans n°1, 2004.

IREM de Poitiers

– *Les Chantiers du chaos : Hasard ; Probabilité ; Modélisation ; Prédictibilité ; Déterminisme*, D. GAUD, J. GUICHARD, L-M. BONNEVAL, J. JACQUESSON, Th. LE GALLIOT, S. PARPAY, C. BLOCH, J. GACOUGNOLLE, C. CHRÉTIEN, 1998.

IREM de Rouen

– *Probabilités à bâton rompus*, J.-C. JOVET ; D. POMMIER, 1998.

IREM de Strasbourg

– *Enseigner les probabilités en classe de première (programmes 1991),* C. DUPUIS et al., 2000.

IREM de Toulouse

– *Les statistiques dans les nouveaux programmes de collège*, P. POMBOURCQ, 1999.


8
2
5
3

Bibliographie structurée

I - Manuels universitaires et monographies sur la statistique et les probabilités

APMEP (1980). *Analyse des données*, Publications de l'APMEP, tome 1, **28** et tome 2, **40**, Paris.

BATES, D. M., CHAMBERS, J.-M. (1992). Nonlinear models. In *Statistical Models*, Chapter 10, J. M. Chambers and T. J. Hastie eds, Wadsworth & Brooks/Cole.

BATES, D. M., WATTS D. G. (1988). *Nonlinear regression analysis and its applications*, Wiley, New York.

BENZÉCRI, J.-P. (1973). *Analyse des donnée,*. Tome 1 : *la taxinomie*, tome 2 : *Analyse des correspondances*, Dunod, Paris.

BOULEAU, N. (1986). *Probabilités de l'ingénieur : variables aléatoires et simulation*, Hermann, Paris.

BOULEAU, N. (1999). *Philosophies des mathématiques et de la modélisation, du chercheur à l'ingénieur*, L'Harmattan, Paris.

BOUROCHE, J.-M., SAPORTA G. (1980). *L'analyse des données*, PUF, coll. Que sais-je, **1854**, Paris.

BOURSIN, J.-L. (1991). *Comprendre la statistique descriptive*, Armand Colin, Paris.

CARNEC, H., SEROUX, R., DAGOURY, J.-M., THOMAS M., (2000). *Itinéraires en statistiques et probabilités*, Ellipses, Paris.

CHAITIN, G. (1999). Les suites aléatoires. *Pour la Science*, dossier hors série : *Le Hasard*, 1996, ou *La Recherche*, dossier *L'univers des nombres*, 1999.

CHAMBERS, J. M., CLEVELAND, W. S., KLEINER, B., TUKEY, P. (1983). *Graphical Methods for Data Analysis*, Wadsworth Belmont, Cal.

CHAUVAT, G., RÉAU, J.-P. (1992) *Statistiques descriptives, Exercices et corrigés*, Armand Colin, Collection Cursus, Paris.

CHAUVAT, G., RÉAU, J.-P. (1995) *Statistique descriptive*, Hachette, coll. Les Fondamentaux, Paris.

CIBOIS, P. (1983). *L'analyse factorielle*, PUF, coll. Que sais-je ?, Paris.

CLUZEL, R., VISSIO, P. et CHARTIER, F. (1966). *Mathématiques et Statistique*, 1ère D. Delagrave, Paris.

DACUNHA-CASTELLE, D. (1996). *Chemins de l'aléatoire*, Flammarion, Paris.

DAGNELIE, P. (1998). *Statistique, théorique et appliquée* (Tomes 1 et 2), De Boeck & Larcier, Paris et Bruxelles.

De FINETTI, B. (1974). *Theory of probability*, Wiley, New York.

DELAHAYE, J.-P. (1997). *Le fascinant nombre* π, Paris, Belin, coll. Pour la Science.
DELAHAYE, J.-P. (1999a). *Logique, informatique et paradoxes*, Belin, coll. Pour la Science, 3ème édition, Paris.
DELAHAYE, J.-P. (1999b). *Information, complexité et hasard*, 2ème édition revue, Hermès, coll. Science Publications, Paris.
DODGE, Y. (1993). *Statistique, Dictionnaire encyclopédique*, Dunod, Paris.
DRESS, F. (1997). *Probabilités, Statistique, rappels de cours, questions de réflexion, exercices d'entraînement*, Dunod, Paris.
DRESS, F., MENDÈS-FRANCE, M. (2001). La suite des puissances de 3/2, *La Recherche* **346**, pp. 34-37.
DROESBEKE, J.-J. (1997). *Eléments de statistique*, 3ème édition, Ellipses, Paris.
ENGEL, A. (1990). *Les certitudes du hasard*, Aléas, Lyon.
FELLER, W. (1950). *An Introduction to Probability Theory and its Applications*, 2 vol., John Wiley & Sons, 3ème éd. : 1968, 4ème éd. : 1971, New York.
FINE, T. L. (1971). *Theories of probability. An examination of foundations*, Academic Press, London.
FOATA, D., FUCHS, A. (1998) *Calcul des probabilités*, Dunod, Paris.
FOURGEAUD, C. & FUCHS, A. (1967). *Statistique*, Dunod, 2ème éd., Paris.
GLAYMANN, M. (1976) : L'enchevêtrement des chiffres d'une table de chiffres aléatoires. *Hasardons-nous* (pp. 125-137), APMEP, brochure **17**, Paris.
HAMMERSLEY, J.-M. et HANDSCOMB, D. C. (1967). *Les méthodes de Monte-Carlo* (traduit de l'anglais), Dunod, Paris.
HARTHONG, J. (1996). *Probabilités & statistiques. De l'intuition aux applications*, Diderot Editeur, coll. Arts et sciences, Paris.
HENNEQUIN, P.-L. (1976). Quelques remarques sur l'article précédent. *Hasardons-nous* (pp. 139-144), APMEP brochure **17**, Paris.
HENNEQUIN, P.-L. (1981). Schéma de Bernoulli et planchettes à clous. *Bulletin de l'APMEP* (pp. 435-441), **329**.
IREM de Strasbourg (1983). *Mathématiques, Terminales C et E, analyse et statistiques*, Librairie Istra, Strasbourg.
ISRAEL, G. (1996). *La mathématisation du réel*, Editions du Seuil, Paris.
JACQUARD, A. (1974). *Les probabilités*, PUF, coll. Que sais-je ? **1571**, Paris.
JANVIER, M. (2001). Les nombres pseudo-aléatoires. *Des statistiques à la pensée statistique*, (pp. 165-187), IREM de Montpellier.
JOLIVET, E. (1983). Introduction aux modèles mathématiques en biologie, I.N.R.A., *Actualités scientifiques et agronomiques*, Masson, Paris.
KAUFFMANN, P. (1994). *Statistique, Information, Estimation, Tests*, Dunod, Paris.
KENDALL, M. G. & STUART, A. (1966). *The advanced theory of statistics* (3 vol.), C. Griffin & Co, London.
LANNUZEL, B. (1999). *Probabilités et statistique, Cours et exercices corrigés*, Dunod, Paris.

LEBART, L., MORINEAU, A., PIRON, M., (1995). *Statistique exploratoire multidimensionnelle*, Dunod, Paris. 2e éd. 1998.

LEBRETON, J.-D., MILLIER, C. & all. (1982). *Modèles dynamiques et déterministes en biologie*, Masson, Paris.

L'ÉCUYER, P. (1988). Efficient and portable combined random number Generators, *Communications of the ACM* **31**-**5**, (pp. 742-749 + 774).

LUKACS, E. (1970). *Characteristic Functions*, C. Griffin, London.

MAISTROV, L. E. (1974). *Probability Theory : A Historical Sketch*, S. Kotz, trad. et éd., New York et Londres, Academic Press.

MARQUARDT, D. W. (1963). An algorithm for least square estimation of non linear parameters, *S.I.A.M.J.*, **11**, pp. 431-441.

MARTIN-LÖF, P. (1966). The definition of random sequences. *Information and Control*, **9**, 602-619.

MÉTIVIER, M. (1972). *Notions fondamentales de la théorie des probabilités*, Dunod, 2ème éd., Paris.

MOLK, J. (1992). *Encyclopédie des Sciences Mathématiques pures et appliquées*, Tome I, Arithmétique et Algèbre, Vol. 4. Calcul des probabilités. Théorie des erreurs. Applications diverses, Gauthier-Villars et Teubner, 1904-1916, Paris, Réédité par J. Gabay, Paris.

PIAGET, J., INHELDER, B. (1951). *La genèse de l'idée de hasard chez l'enfant*, PUF, Paris.

PIEDNOIR, J.-L. (1977). *Statistiques non paramétriques*, Cethedec, Paris.

Rand Corporation (1955). *A million Random Digits with 100,000 Normal Deviates*, The Free Press, Illinois.

RENYI, A. (1966). *Calcul des probabilités*, Dunod, Paris, Rééd. Jacques Gabay, Paris, 1992.

ROBERT, C. (1989). *Analyse descriptive multivariée*, Flammarion médecine-sciences, Paris.

ROBERT, C. (1995). *L'empereur et la girafe - Leçons élémentaires de statistiques*, Diderot Editeur, Paris.

ROHLF, F. J. & SOKAL, R. (1969). *Statistical Tables*, W.H. Freeman & Co, San Francisco.

RUELLE, D., (1991). *Hasard et chaos*, Odile Jacob, Paris.

SAPORTA, G. (1990). *Probabilités, Analyse des données et Statistique*, Technip, Paris.

SCHLACTHER, D. (1986). *De l'analyse à la prévision*, Ellipses, Collection Statistique pour les sciences économiques et sociales, Paris.

SCHWARTZ, D. (1994). *Le jeu de la science et du hasard*, Flammarion, Paris.

SPIEGEL M. R. (1993). *Théorie et applications de la statistique*, Mc-Graw-Hill, 2ème édition, Paris.

SPIEGEL, M. R. (1993). *Statistique, Cours et problèmes*, McGraw Hill (Ediscience international), New York. Ed. française par A. ERGAS et J.-F. MARCOTORCHINO, Série Schaum, Paris.

TASSI, P. (1985). *Méthodes statistiques*, Economica, Paris.

TOMASSONE, R., AUDRAIN, S., LESQUOY-de-TURCKHEIM, E., MILLIER C. (1992). *La régression : nouveaux regards sur une ancienne méthode statistique*. Masson, Paris.

TOMASSONE, R., ROUX C., (1973). Ajustements non-linéaires (HAUSS 59), *Note interne du Laboratoire de Biométrie du C.N.R.S.*

TUKEY, J. W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading, Mass.

VENTSEL, H. (1973). *Théorie des probabilités*, Mir, Moscou.

VEYSSEYRE R. (2000). *Statistique et probabilités pour l'ingénieur*, Dunod, Paris.

WONNACOTT, T. H. & WONNACOTT, R. J. (1995). *Statistique*, Economica.

II - Travaux sur l'enseignement de la statistique et des probabilités, études didactiques

ARTIGUE, M., PARZYSZ, B. (2003). Causalités et dépendances : quelle place dans les recherches en didactique des mathématiques, *Enquête sur le concept de causalité*, VIENNOT (éd.), Presses Universitaires de France, Coll. Science, Histoire et Société, (pp. 123-151), Paris.

BADIZÉ M., JACQUES A., PETITPAS M. & PICHARD J.-F. (1996). *Le jeu du franc-carreau - une activité probabiliste au Collège*, IREM de Rouen.

BATANERO, C. (2001). *Didactica de la Estadistica*, Departamento de didactica de la Matematica, Universidad de Granada, Consultable en ligne : http://www.ugr.es/local/batanero.

BORDIER, J. (1991). *Un modèle didactique utilisant la simulation sur ordinateur, pour l'enseignement de la probabilité*, Thèse de doctorat, Université Paris-7, Paris.

BOROVCNIK, M. & KAPADIA, R. (1991). *Chance encounters : probability in education*, Kluwer Academic Publishers, Dordrecht.

BOROVCNIK, M. and PEARD, R. (1996). Probability. In *International Handbook of Mathematics Education*, chapter 7, A. J. Bishop et all. Eds, pp. 239-287, Kluwer Academic Publishers, Dordrecht.

CHRÉTIEN, C., GAUD, D. (1998). Qu'est-ce que le hasard ? Comment le mathématiser ? In *Repères IREM*, **32**, pp. 81-110, Topiques Editions, Metz.

Commission de Réflexion sur l'Enseignement des Mathématiques (2002). Statistiques et Probabilités. Dans J.-P. Kahane (éd.) *L'enseignement des sciences mathématiques* (pp. 51-86), rapport au ministre de l'Education nationale, Odile Jacob, Paris.

Commission Inter IREM Histoire et Epistémologie des Mathématiques, Actes de la 6ème U. E., 1995, IREM de Besançon.

Commission Inter IREM Lycées Technologiques (1998). *Simulation d'expériences aléatoires. Une expérience du hasard de la première au BTS sur calculatrice et ordinateur*, B. Verlant (éd.), brochure **93**, IREM de Paris Nord.

Commission Inter IREM Lycées Technologiques (2000). *Simulation et statistique en Seconde*, B. Verlant (éd.), brochure **102**, IREM de Paris Nord.

Commission Inter IREM Lycées Technologiques (2001). *Enseigner la statistique au lycée : des enjeux aux méthodes*, J.-L. Piednoir & P. Dutarte (éds.), brochure **112**, IREM de Paris Nord.

Commission Inter IREM Lycées Technologiques (2002). *La statistique inférentielle en quatre séances*, B. Verlant (éd.), brochure **118**, IREM de Paris Nord.

Commission Inter IREM Lycées Technologiques (2003). *Le nouveau programme de statistique et probabilités au lycée*, B. Verlant (éd.), brochure **124**, IREM de Paris Nord.

Commission Inter-IREM Statistique et Probabilités (1993-1995). *Actes des Universités d'été de Statistiques Inférentielles*, La Rochelle : 1-5 septembre 1992 et Rouen : 29 août-2 septembre 1994. J.-F. Pichard (éd.), IREM de Rouen.

Commission Inter-IREM Statistique et Probabilités (1997). *Actes de l'Université d'été de Probabilités*, Metz 26-31 août 1996, J.-F. Pichard (éd.), IREM de Lorraine.

Commission Inter-IREM Statistique et Probabilités (1997). *Enseigner les probabilités au lycée*, B. Chaput et M. Henry (éds.), IREM de Reims.

Commission Inter-IREM Statistique et Probabilités (2001). *Autour de la modélisation en probabilités*, M. Henry (éd.), Presses Universitaires Franc-Comtoises, coll. Didactiques, Besançon.

Commission Inter-IREM Statistique et Probabilités (2003). *Probabilités au lycée*, B. Chaput (éd.), APMEP, brochure **143**, Paris.

COURIVAUD, J. (1991). Le traitement graphique des images de géométrie, *Repères-IREM* **4**, (pp. 5-20), Topiques Editions, Metz.

COUTIÑHO C. (2001). *Introduction aux situations aléatoires dès le collège : de la modélisation à la simulation d'expériences de Bernoulli dans l'environnement informatique Cabri-Géomètre 2*, Thèse de doctorat, Université Joseph Fourier, Grenoble I.

DUTARTE, P., KERN, C. & all. (1998). *La statistique inférentielle en quatre séances*, Commission Inter-IREM Lycées technologiques, brochure **118**, B. Verlant (éd.), IREM de Paris-Nord.

DUTARTE, P., KERN, C. & all. (2000). *Simulation et statistique en seconde*. Commission Inter-IREM Lycées technologiques, broch. **102**, B. Verlant (éd.), IREM de Paris-Nord.

DUTARTE, P., KERN, C. & all. (2003). *Le nouveau programme de statistique et probabilités au lycée*, Commission Inter-IREM Lycées technologiques, brochure **124**, B. Verlant (éd.), IREM de Paris-Nord.

DUTARTE, P., KERN, C. NOUGUÈS, M.-F., SAINT-PIERRE, G., VERLANT, B. (1998). *Simulation d'expériences aléatoires. une expérience du hasard de la première au BTS sur calculatrice et ordinateur*, Commission Inter-IREM Lycées technologiques, brochure **93**, IREM de Paris-Nord.

FISCHBEIN, E. (1975). *The intuitive sources of probabilistic thinking in children*, Reidel, Dordrecht.

G.E.P.S. (2001). *Accompagnement des programmes de lycée, Mathématiques, rentrée 2002*, Ministère de la Jeunesse, de l'Education et de la Recherche, CNDP, Paris.

G.R.E.S. (Groupe de Réflexion sur l'Enseignement de la Statistique) (1995-2000). Bulletins du G.R.E.S., Ministère de l'Agriculture et de la Pêche, ENFA, Toulouse-Auzeville.

GAL, I. & GARFIELD, J. B. (eds, 1997). *Assessement Challenge in Statistics Education*, International Statistical Institute (ISI) & International Association for Statistical Education (IASE).

GIRARD, J. C. (1998). La médiane, pour quoi faire ? Un exemple d'utilisation : les boîtes de dispersion, *Enseigner la statistique du CM à la seconde. Pourquoi ? Comment ?* IREM de Lyon.

GIRARD, J. C. (2001). Un exemple de confusion modèle-réalité. *Autour de la modélisation en probabilités*, Commission inter-IREM Statistique et Probabilités, ouv. cité, (pp. 145-148).

GIRARD, J. C. (2003). Difficultés et obstacles dans l'enseignement des probabilités. *Probabilités au lycée*, brochure APMEP **143** (pp. 35-48), Paris.

GIRARD, J. C., HENRY, M., PARSYSZ, B., PICHARD, J. F. (2001). Quelle place pour l'aléatoire au collège ? *Repères-IREM* **42**, (pp. 27-43), Topiques Editions, Metz.

GIRARD, J. C. & PARZYSZ, B. (1998). De la modélisation en mathématiques, In *Bulletin APMEP* **418**, pp. 573-582.

GIRARD, J. C. (1998). A bas la moyenne ! *Repères-IREM* **33** (pp. 97-114), Topiques Editions, Metz.

GIRARD, J. C. (1999). Le professeur de mathématiques doit-il enseigner la modélisation ?, *Repères-IREM* **36** (pp. 7-14), Topiques Editions, Metz.

GRANGÉ, J.-P. (2003). Arbres et tableaux en probabilités conditionnelles, *Probabilités au lycée*, brochure APMEP **143** (pp. 91-124), Paris.

GROS, D., (2001). Une enquête statistique au service de la proportionnalité. Dans *Repères-IREM*, n°**44**, (pp. 69-80), Topiques Editions, Metz.

HENRY, M. (1994). *L'enseignement des probabilités - perspectives historiques, épistémologiques et didactiques*, IREM de Besançon.

HENRY, M. (1999). L'introduction des probabilités au lycée : un processus de modélisation comparable à celui de la géométrie, *Repères-IREM* **36** (pp. 15-34), Topiques Editions, Metz.

HENRY, M. (2001). Notion de modèle et modélisation dans l'enseignement, *Autour de la modélisation en probabilités*, Commission inter-IREM Statistique et Probabilités, ouv. cité.

LAHANIER-REUTER, D. (1998). *Etude de conceptions du hasard : approche épistémologique, didactique et expérimentale en milieu universitaire*, Thèse de Doctorat, Rennes : Université de Rennes I.

Ministère de l'Education Nationale, Direction de l'Evaluation et de la Prospective (1993). *Repères et références statistiques sur les enseignements et la formation 1991-1992*. *Repères et références statistiques sur les enseignements et la formation 2001*, Paris.

PARZYSZ B. (1997). L'enseignement de la statistique et des probabilités dans l'enseignement secondaire, d'hier à aujourd'hui. *Enseigner les probabilités au lycée* (pp. 17-38), Commission Inter-IREM Statistique probabilités, IREM de Reims.

PARZYSZ B. (2003). L'enseignement de la Statistique et des probabilités en France : évolution au cours d'une carrière d'enseignant (période 1965-2002). *Probabilités au lycée*, (pp. 9-34), Commission Inter-IREM Statistique et Probabilités, brochure APMEP **143**, Paris.

PARZYSZ, B. (1980). Les mots pour le dire. Sur le vocabulaire des dénombrements, *Groupe français-mathématiques, Volume 2*, (pp. 133-38), IREM de l'Université Paris-7.

PARZYSZ, B. (1993). Des statistiques aux probabilités : exploitons les arbres, *Repères-IREM* **10** (pp. 91-104), Topiques Editions, Metz.

PICHARD, J.-F. (1998). Approche épistémologique et diverses conceptions de la probabilité, *Repères-IREM* **32** (pp. 5-24), Topiques Editions, Metz.

PIEDNOIR, J.-L. & DUTARTE, P. (2001). *Enseigner la statistique au lycée : des enjeux aux méthodes*, Commission Inter-IREM Lycées technologiques, brochure **112**, IREM Paris-Nord.

RAYMONDAUD, H. (2003). Modèles d'urnes pour introduire et simuler quelques lois discrètes, *Probabilités au lycée* (pp. 51-74), Commission Inter-IREM Statistique et Probabilités, brochure APMEP **143**, Paris.

STEINBRING, H. (1991). The Theoretical Nature of Probability in the Classroom. In R. Kapadia & M. Borovcnik (eds.), *Chance Encounters: Probability in Education*, coll. Mathematics Education Library (pp. 135-167), Kluwer Academic Publishers, Dordrecht.

STEINBRING, Heinz (1986). L'indépendance stochastique. Un exemple de renversement du contenu intuitif d'un concept et de sa définition mathématique formelle, *Recherches en Didactique des Mathématiques* vol. 7 n° **3**, (pp. 5-50), La pensée sauvage, Grenoble.

THIÉNARD J-C. (1993). *À propos de l'enseignement du calcul des probabilités*, IREM de Poitiers.

ZAKI, M. (1990). *Traitements de Problèmes de Probabilités en Situation de Simulation*, Thèse de Doctorat de l'Université Louis Pasteur, Strasbourg.

III - Travaux sur l'histoire de la statistique et des probabilités

BELLHOUSE, D. R. (2000). De Vetula: a medieval manuscript containing probability calculations, *International Statistical Review*, 68 (2), 123-136.

BENZÉCRI, J.-P. (1976). Histoire et préhistoire de l'analyse des données. *Cahiers de l'analyse des données*, t. I,1, 2, 3, 4 ; t. II,1, Dunod, Paris.

BOYÉ, A., LEFORT, X. (1996). De Cassini à Gauss : du calcul d'erreurs aux probabilités, *Actes de la 6e Université d'été sur l'histoire des mathématiques*, 1995, IREM de Besançon.

BRU B. (1981). Petite Histoire du Calcul des Probabilités, In *Fragments d'Histoire des Mathématiques*, Brochure APMEP n° 41, pp. 141-158, Paris.

BRU, B. (1988). Laplace et la critique probabiliste des mesures géodésiques, *La figure de la Terre du XVIIIe siècle à l'ère spatiale*, H. Lacombe et P. Costabel éds., Gauthiers-Villars, Paris.

CLÉRO, J.-P. (1988). *Cahiers d'Histoire et de philosophie des Sciences*, n° 18, 1988.

Commission Inter-IREM Histoire et Epistémologie des mathématiques (2004). *Histoire de probabilités et de statistiques*, Actes du 14ème colloque, Orléans 2002, E. Barbin & J.-P. Lamarche (coords.), Ellipse, Paris.

COURTEBRAS, B. (2001). Sur quelques conceptions du hasard, *Autour de la modélisation en probabilités*, Commission inter-IREM Statistique et Probabilités, ouv. cité, (pp. 145-148).

DROESBEKE, J.-J., TASSI, P. (1990). *Histoire de la Statistique*, PUF, coll. Que sais-je ? **2527**, Paris.

HACKING, I. (2002). *L'émergence de la probabilité*, (1ère éd.: The emergence of probability, Cambridge, MA: Cambridge University Press, 1975), Le Seuil, Paris.

KENDALL, M. G. & PLACKET, R. L. eds. (1977). *Studies in the history of statistics and probability* (2 vol.), C. Griffin & Co, London.

MAISTROV, L. E. *Probability Theory : A Historical Szketch*, S. Kotz, trad. et éd., New York et Londres, Academic Press, 1974.

MARTIN, T. (1996). *Probabilités et critique philosophique selon Cournot*, Coll. Mathesis, Blay & Sinaceur (dir.), Librairie Philosophique J. Vrin, Paris.

MEUSNIER, N. (1989). Argumentation et démonstration : A quoi sert la démonstration de la "Loi des grands nombres" de Jacques Bernoulli (1654-1705), *La démonstration mathématique dans l'histoire*, 7ème colloque Epistémologie et Histoire des mathématiques, IREM de Besançon.

PEARSON, E. S. and KENDALL, M. G. eds. (1970). *Studies in the history of statistics and probability*, vol. 1, C. Griffin & Co, London.

PEARSON, K. (1978). *The History of Statistics in the 17th & 18th Centuries*, ed. E. S. Pearson, Ch. Griffin & Co, London.

PICHARD, J.-F. (2001). Les probabilités au tournant du XVIIIe siècle, *Enseigner les probabilités au lycée*, Commission Inter-IREM Statistique et Probabilités, IREM de Reims, 1997, Réed. dans *Autour de la modélisation en probabilités*, ouv. cité.

STIGLER, S. M. (1986). *The History of Statistics, The Measurement of Incertainty before 1900*, Belknap Press of Harvard University Press, Cambridge, Massachusetts.

TODHUNTER, I. (1865). *A History of the Mathematical Theory of Probability*. Cambridge, 1865. Rééd. Chelsea, New York, 1965.


7
8
12
5

Œuvres anciennes citées

AGNESI, M.G. (1748). *Instituzioni Analitiche...* (Milan). Voir l'article "Maria Gaetana Agnesi", *L'Ouvert* **96**, sept. 1999, IREM de Strasbourg.

ARBUTHNOT, John (1710). An argument for Divine Providence, taken from the constant regularity observed in the Birthd of both sexes, *Philosophical Transactions*, **27**, 186-90. Reprod. dans KENDALL, M. et PLACKETT, R.L. (éds) (1977). *Studies in the History of Statistics and Probability*, Vol. II. Ch. Griffin & Co, London.

ARCHIMÈDE (IIe siècle av. J.-C.). *Oeuvres*, tome II, traduit par C. Mugler, éd. Les Belles Lettres, Paris, 1971.

BAYES, Thomas (1763). An Essay towards solving a Problem in the Doctrine of Chances. *Philosophical Transactions*, London. « Thomas BAYES, Essai en vue de résoudre un problème de la doctrine des chances », traduit par Cléro, J.P., *Cahiers d'histoire et de philosophie des sciences* **18**, 1988.

BERNOULLI, Daniel (1738). "Specimen theoriae novae de mensura sortis". *Commentarii academiae scientarum imperialis Petropolitanae*, t. **5** pour 1730-31. Traduit dans "Exposition of a new theory on the measurement of risk". *Econometrica*, Vol. **22,** n°**1**, 1954.

BERNOULLI, Daniel (1778). Dijudicatio maxime probabilis plurium observationum discrepantium ... *Novi Comm. Acad. Sc. Imp. Petrop* (pour 1777). Traduit "*The most probable choice...*" dans Pearson et Kendall, *Studies I*.

BERNOULLI, Jakob (1713). *Ars Conjectandi,* 4ème partie. Traduit du latin par Norbert MEUSNIER dans *Jacques Bernoulli et l'Ars conjectandi*, IREM de Rouen, 1987.

BERTILLON, Adolphe (1876). Moyenne. *Dictionnaire encyclopédique des sciences médicales*, 2e série. Paris

BERTRAND, Joseph (1822-1900) : *Calcul des probabilités*, Gauthier-Villars, Paris, 1889 (2e éd. 1907).

BOREL, Emile (1938), *Valeur pratique et philosophie des Probabilités*, 2ème édition, 1952.

BUFFON, Georges Louis LECLERC de (1777). *Essai d'arithmétique morale*, dans *Œuvres Complètes*, tome **12**, Ed. Garnier, 1855. Reproduit dans *Un autre Buffon*, Binet, J. L. et Roger, J., Hermann, 1977.

CARDANO, Gerolamo (vers 1560), *Liber de Ludo Aleae*, publié dans *Opera*, Lyon, 1663 ; trad. *The Book on Games of Chances*, Holt, Rinehart and Winston. New-York, 1961.

CHAMPERNOWNE, D.G. (1933). "The construction of decimals normal in the scale of ten". *J. London Math. Soc*. **8**, 254-60.

CONDORCET Jean Nicolas de Caritat, Marquis de (1767-1789). *Arithmétique politique, Textes rares ou inédits*, Ed. B. Bru et P. Crepel, INED, 1994.

CONDORCET Jean Nicolas de Caritat, Marquis de. *Elémens du calcul des probabilités et son application aux jeux de hasard, à la loterie et aux jugemens des hommes. Avec un discours sur les avantages des mathématiques sociales,* A Paris, chez Royez, libraire, An XIII.

CRAMÉR, Harald (1937). *Random variables and probability distributions*, Cambridge University Press (2e éd.), 1961.

DE MOIVRE, Abraham (1718). *The Doctrine of Chances* (1e éd., 1718 ; 2e éd., 1738 ; 3e éd., 1756). Reproduit : 1967, 2000, New York, Chelsea.

EULER, Leonhard (1748-1773). *Œuvres (Leonhardi Euleri Opera Omni)* série I – Opera mathematica, Leipzig et Berlin, Teubner, vols. **17** à **21**, 1912-1932.

FOURIER, Joseph (1822). *Théorie Analytique de la Chaleur* (réédition, J. Gabay, 1988).

FRÉCHET, Maurice (1937). *Généralités sur les Probabilités. Variables aléatoires.* Gauthier-Villars, Paris.

GALILEI, Galileo (vers 1620). Considerazione sopra il Giuco dei Dadi, *Opere de Galileo Galilei*, Firenze, 1855, t. **xiv**, p. 293-296 ; 1ère publication des *Opera*, Florence, 1718.

GALTON, Francis (1889). *Natural Inheritance*, Macmillan, London ;

GAUSS, Karl Friedrich (1809). (1821-26) *Méthode des moindres carrés. Mémoires sur la combinaison des observations*, Trad. J. Bertrand, Paris, 1855. Reproduit dans *Friedrich Gauss. Méthode des moindres carrés,* IREM de Paris VII, 1996.

HUYGENS, Christiaan (1657). *De ratiociniis in Ludo aleae* ; trad. "Du calcul dans les jeux de hasard" in tome 14, *Œuvres complètes* **22,** vol. 1888-1950, La Haye.

KENDALL Maurice G. and BABINGTON SMITH B.(1939). *Tables of random sampling numbers*. Tracts for computers, **24**, Cambridge University Press.

KRAMP, Christian (1799). *Analyse des réfractions astronomiques et terrestres*, Strasbourg et Leipzig.

LAGRANGE, Joseph Louis. (1773). Mémoire sur l'utilité de la méthode de prendre le milieu entre les résultats de plusieurs observation. *Misc. Taurinensia* (1770/73), publié en 1776. *Œuvres*, tome 2, Paris, 1868.

LAPLACE, Pierre Simon de (1814). *Essai philosophique sur les probabilités* (5ème édition, 1825), préface de René THOM, postface de B. BRU, Editions Bourgois, 1986.

LAPLACE, Pierre Simon de. (1774), Mémoire sur la probabilité des causes par les événements, *Oeuvres Complètes*, tome 8, Gauthier-Villars, 1889.

LAPLACE, Pierre Simon de. (1779-1825). *Mécanique Céleste*, vol. I et II, 1799 ; Vol. III, 1803 ; vol. IV, 1805, vol. V, 1825.

LAPLACE, Pierre Simon de. (1785). Mémoire sur les approximations des formules qui sont fonctions de très grands nombres. *Mémoires de l'Académie royale des Sciences de Paris*, année 1782 ; 1785.

LAPLACE, Pierre Simon de. (1811).Mémoire sur les intégrales définies, et leur application aux probabilités, et spécialement à la recherche du milieu qu'il faut choisir entre les résultats des observations, *Œuvres de Laplace*, tome XII, Paris.

LAPLACE, Pierre-Simon de (1812) : *Théorie analytique des probabilités*. Tome VII des *Œuvres de Laplace*. Paris, Imprimerie royale (1847). Réédition Jacques Gabay 1995.

LEGENDRE, Adrien Marie (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*, Paris.

LÉVY, Paul (1923). *Calcul des probabilités*, Gauthier-Villars, Paris.

LÉVY, Paul. (1937). *Théorie de l'addition des variables aléatoires*. Gauthier-Villars, Paris.

MONTMORT, Pierre Rémond de (1708). *Essay d'analyse sur les jeux de hazard* ; 2e édition, 1713.

PEARSON, Karl (1914-1930). *The Life, Letters and Labours of Francis Galton* (3 vols.) Cambridge, Cambridge University Press.

PEARSON, Karl (1920). "Note on the History of Correlation", *Biometrika*, **13**, 25-45 ; reproduit dans Pearson et Kendall, *Studies*..., vol. I.

PEARSON, Karl (1934). *Tables of the Incomplete Beta-function*. 2e éd. 1968, Biometrika Trustees

PEARSON, Karl. Contributions to the mathematical theory of evolution, *Philosophical Transactions of the Royal Society of London* (1893-95), puis Mathematical contributions to the theory of evolution, *Philosophical Transactions of the Royal Society of London* (1896-1905).

POLYÁ, György (1920). Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem, *Math Zeitschrift*, tome 8.

QUETELET, Adolphe (1827). "Recherches sur la population, les naissances, les décès, ... dans le royaume des Pays-Bas". *Nouveaux mémoires de l'Académie des sciences et belles-lettres de Bruxelles*.

QUETELET, Adolphe (1835). *Sur l'homme et le développement de ses facultés, ou essai de physique sociale*. Paris.

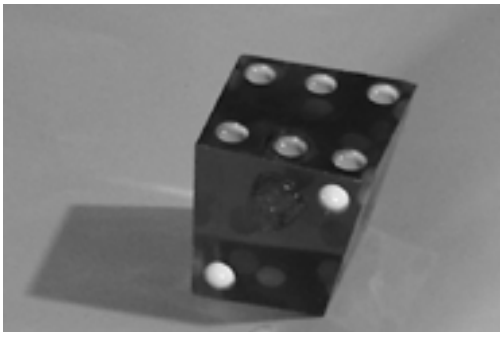
QUETELET, Adolphe (1846). *Lettres à S.A.R. le Duc Régnant de Saxe-Cobourg et Gotha sur la théorie des probabilités, ...* Bruxelles.

STUDENT (William Sealy GOSSET, 1908), Probable error of a correlation coefficient. *Biometrika*, **6**, 302.

STUDENT (William Sealy GOSSET, 1927). Errors of routine analysis, *Biometrika* **19**.

TODHUNTER, Isaac. *A History of the Mathematical Theory of Probability*. Cambridge, 1865. Rééd. Chelsea, New York, 1965.

WILSON, E.-B. (1923). First and Second Laws of Error. *Quaterly publications of the Amer. Stat. Assoc.*



Index terminologique

Index des noms des personnes citées

(hors bibliographie)

Sommaire du volume 2 (en préparation)

Activités statistiques pour la classe

Première partie : Simulation

Deuxième partie : Sondages

Méthodes de sondages : historique et description
Jean-François PICHARD

Pour une lecture critique des sondages
Jean Claude GIRARD

Échantillonnage et fourchettes de sondage
Brigitte CHAPUT, Michel HENRY

TP sur les sondages
Brigitte CHAPUT

Troisième partie : Traitements d'enquêtes

Séries univariées quantitatives
Hubert RAYMONDAUD

Séries bivariées quantitatives
Hubert RAYMONDAUD

Traitement d'une enquête qualitative
Hubert RAYMONDAUD

Les auteurs du volume 1

(adresses professionnelles)

Louis-Marie BONNEVAL
IREM de POITIERS
Université de Poitiers
40, av. du Recteur Pineau
86022 POITIERS CEDEX

Brigitte CHAPUT
IREM de TOULOUSE
ENFA, B.P. 22687
31326 CASTANET-TOLOSAN CEDEX

Annette CORPART
IREM de CLERMONT-FERRAND
Lycée Jean Zay
63300 THIERS

Jean-Claude GIRARD
IREM de LYON
IUFM, centre local de Saint-Étienne
90, rue Richelandière
42100 SAINT-ETIENNE

Michel HENRY
IREM de BESANÇON
Université de Franche-Comté, IREM
UFR des Sciences et des Techniques
16 route de Gray
25030 BESANÇON CEDEX

Stéphan MANGANELLI
ENSEIGNEMENT AGRICOLE
Lycée d'Enseignement Général et Technique Agricole Louis GIRAUD
Hameau de Serres
84200 CARPENTRAS

Bernard PARZYSZ
IREM de PARIS 7
Equipe DIDIREM (université Paris-7)
IUFM d'Orléans-Tours
72, faubourg de Bourgogne
45044 ORLEANS CEDEX 1

Jean-François PICHARD
IREM de ROUEN
Université de Rouen, IREM
UFR des Sciences et des Techniques
Boulevard de Broglie
76821 MONT-SAINT-AIGNAN CEDEX

Hubert RAYMONDAUD
ENSEIGNEMENT AGRICOLE
Lycée d'Enseignement Général et Technique Agricole Louis GIRAUD
Hameau de Serres
84200 CARPENTRAS

Site de la Commission Inter IREM *Statistique et probabilités* :

http://www.univ-lyon1.fr/IREM/CII-Stat/accueil_ciistat.html

Des structures efficaces et démocratiques
sur la base du *bénévolat intégral*,
un ***Comité*** et un ***Bureau National***
des ***Commissions Nationales***
organisées par niveaux ou par thèmes,
lieux d'information et d'échanges, **ouvertes à tous**.

LES JOURNEES NATIONALES
temps fort de l'**APMEP**,
sur 3 à 4 jours, déplaçant près de 1000 enseignants
de mathématiques
dans une Régionale différente chaque année,
sur un thème déterminé,
avec des *conférenciers renommés*, des débats,
de multiples *ateliers*, des *expositions*, ...

26 REGIONALES *organes de liaison*
avec les autorités pédagogiques
et administratives de la Région, très souvent
impliquées au niveau des IREM et des IUFM
relais essentiel entre le National et les adhérents
avec souvent leurs propres bulletins et leurs publications.

6 numéros du
Bulletin à Grande Vitesse ("BGV"),
environ 80 pages par an, en A4.
Véritable ***journal d'information***
pour cerner rapidement
l'actualité mathématique et associative.

Un Serveur WEB
accessible par **Internet**
http://www.apmep.asso.fr

EVAPM, *observatoire*
*d'***EVA***luation par l'***APMEP**
de l'impact des programmes
de mathématiques
au Collège et aux Lycées.

L'

APMEP

c'est

EVAPMIB
base informatisée et mémoire
des évaluations **EVAPM** et autres ...

PUBLIMATH,
banque de données bibliographiques
sur **Internet**,
en association avec les IREM.

Un secrétariat permanent
26 rue Duméril - 75013 PARIS
Tél. 01 43 31 34 05 - Fax : 01 42 17 08 77
Courriel : apmep@apmep.asso.fr

PLOT
(4 numéros par an, en A4,
chacun de 32 pages).
Partager, Lire, Ouvrir, Transmettre
plus spécialement destiné aux "débutants",
lieu d'échanges à de nombreuses rubriques.

6 numéros du
BULLETIN VERT
en 17 x 24, environ 800 pages par an.
Lieu d'échanges avec :
- des **articles de fond** pour continuer à se former
- des **documents "clés en main"** de l'école au lycée,
- des **problèmes** pour entretenir son esprit de recherche
- une **documentation riche** sur des publications récentes,
- un **dossier** par numéro ...

. **Les Brochures ou Cédéroms**
par niveau d'enseignement
et par secteur avec **réduction de 30%**
aux **adhérents** et **abonnés.**
. **Co-éditions** et **codiffusions**
à prix réduit, d'autres ouvrages.

des **Groupes de Travail**
selon les besoins ou sur projets comme :
Jeux, Brochures LP,
Histoire des maths, Textes d'orientation,
Réflexion sur les programmes de collège, ...

Voila pourquoi l'A.P.M.E.P. a besoin de vous ...
Pour renforcer son action, VOTRE ACTION !

Commission Inter-IREM
Statistique et Probabilités

Titre : **Statistique au lycée. Volume 1 : Les outils de la statistique**

Date : Juillet 2005

Auteurs : Louis-Marie BONNEVAL, Brigitte CHAPUS, Annette CORPART, Jean Claude GIRARD, Michel HENRY, Stéphan MANGANELLI, Bernard PARZYSZ, Jean-François PICHARD, Hubert RAYMONDAUD

Préface : Jean-Pierre RAOULT

Coordination et réalisation : Brigitte CHAPUS et Michel HENRY

Après Probabilités au lycée (brochure APMEP n° 143, 2003), la Commission Inter-IREM *Statistique et Probabilités* a entrepris un travail de fond sur l'enseignement de la statistique tel qu'il est conçu dans le cadre des programmes des années 2000 des lycées.
Cette publication *Statistique au lycée* est livrée en deux volumes :

1 - Les outils de la statistique
2 - Activités statistiques pour la classe.

Le présent volume est conçu comme une introduction, un débat et un élargissement autour des questions d'enseignement soulevées par les objectifs et la démarche adoptés dans l'ensemble des programmes de la seconde à la terminale. Le second volume à paraître en 2006, est plus centré sur les questions relatives à l'échantillonnage, aux situations de sondage, ainsi qu'à des exemples de simulation avec ou sans tableur.
On trouvera dans ce volume une bibliographie étoffée sur les travaux, anciens ou actuels, sur les probabilités et la statistique et leur enseignement.