Éléments primaires de la modélisation statistique en traitement de données

Joseph SAINT PIERRE,
Centre Interuniversitaire
de Calcul de Toulouse
118, Route de Narbonne
31062 Toulouse Cedex 04
Tél: 05 61 36 60 21
Fax: 05 61 52 14 58
e-mail: Joseph.Saint-Pierre@cict.fr

Vendredi 27 Février 1998

1 Préambule, retour sur le séminaire précédent.

En guise d'introduction au séminaire de ce jour je voudrais revenir sur certains éléments et certaines questions qui m'ont été posées la dernière fois notamment après l'exposé. Je tiens à rappeler que je suis ingénieur au CICT et non enseignant à l'Université qu'il ne fait pas vraiment partie de mes attributions de me substituer aux enseignants de statistiques. Le CICT n'est pas non plus un centre de formation même si ponctuellement le personnel du CICT intervient sur des formations. Pour plus de précisions sur les missions du CICT, si vous avez accès à Internet je vous conseille de regarder la page WEB du CICT sur http://www.cict.fr. Si des enseignements de statistiques spécifiques pour les étudiants de sciences humaines de l'UTM sont à mettre en place c'est à cette Université d'en décider. Le séminaire du 16 Janvier 1998 n'était pas une formation dont l'ambition aurait été de faire comprendre l'ensemble des méthodes présentées mais plutôt une tentative de sensibilisation aux statistiques. Il aurait été parfaitement illusoire de croire qu'on puisse expliquer le χ^2, l'AFC simple et multiple, la régression logistique et le modèle log linéaire en trois heures. De même il est important de voir qu'un séminaire ne peut pas être le lieu d'apprentissage à l'utilisation des méthodes de manière concrète, il est prévu d'organiser des séances de formation à l'utilisation de SPSS dans le cadre de l'IED mais c'est pour plus tard. Il existe sur Toulouse, à l'université Paul Sabatier, un très gros laboratoire, le Laboratoire de Statistique et Probabilités dans lequel des enseignants chercheurs acceptent de faire, en plus de leur charge de travail habituelle, de la consultation en direction des autres universitaires, il m'arrive assez souvent de rediriger les questions pointues de statistiques vers les personnes de ce laboratoire, mais j'insiste bien sur le fait que laboratoire LSP est une unité de recherche et non de services et consultations statistiques. Pour une description du laboratoire consultez sur Internet la page WEB du LSP http://www-sv.cict.fr/lsp/. Cela permet de comprendre que les statistiques et probabilités constituent un vaste domaine dans lequel il est très difficile de tout connaître, même superficiellement.

2 Tentative de réponse à une question.

Lors de l'exposé du 16 Janvier 1998, il y a eu une question très intéressante, à mon avis, demandant pourquoi s'était posé la question de la différence d'attitude de la justice américaine entre les accusés noirs et les accusés blancs en cas de meurtre. L'actualité depuis cet exposé a montré que cette question était toujours présente dans la société américaine. Dans le Monde du Jeudi 5 Février 1998, suite à l'exécution de Karla Faye Tucker, il y avait des articles sur cinq pages dont la Une consacrées à la peine de mort aux USA, la page deux y était entièrement consacrée avec des chiffres et des cartes. Il y a un encart en gras dont voici un extrait: «Les adversaires de la peine de mort critiquent tout particulièrement l'inégalité avec laquelle elle est appliquée, selon la race et la fortune de l'accusé: alors que les noirs représentent 12% de la population américaine, ils constituent 41% des condamnés à mort. Bien que moins de la moitié des victimes de meurtres soient des blancs 70% des des gens exécutés depuis vingt ans avaient tué des blancs.» La Floride figure en deuxième position après le Texas pour le nombre d'exécutions. Il y a eu le 16 Janvier une question demandant pourquoi un journal américain s'était posé la question de l'équité de traitement entre noirs et blancs par rapport à la peine de mort dans les années 70; en 1998 la question semble encore d'actualité. Je reproche au journal le Monde d'avoir donné les statistiques pour les races mais pas pour la richesse, alors que le texte dit explicitement que l'application de la peine dépend de la fortune de l'accusé.

3 Autre retour, la triangularité.

Autre remarque, sur le contenu, plus délicate, l'exposé était intitulé «Faut-il se méfier de l'utilisation des statistiques en science humaines?», mais l'exemple simple qui était utilisé concernait la peine de mort et se rapportait plus au savoir juridique qu'au domaine strict des sciences humaines. Cela était, en grande partie, intentionnel. En effet le principe essentiel de l'exposé était fondé sur la différence entre une approche statistique à deux variables et une approche à trois variables la troisième variable venant changer le sens de la liaison entre les deux premières. Pour rappel, la peine de mort est plus souvent appliquée aux blancs mais si on considère la couleur de la victime la peine de mort est plus souvent appliquée aux noirs. La position du juridique, de l'économique et du politique est une sorte de variable tierce par rapport à la liaison binaire entre les sciences mathématiques ou physique et la sociologie, l'anthropologie l'ethnologie ou la psychologie. Géographiquement parlant, l'exposé mettait en opposition Le Mirail et Rangueil en passant par le centre ville et l'Arsenal, institutionnellement, opposition entre Toulouse II et Toulouse III en passant par Toulouse I. Il est sûr que cela était plus facile ou moins risqué car j'évitais ainsi le risque de tomber sur le sujet de recherche de quelqu'un se trouvant dans la salle. Mais cela a aussi un rapport avec l'histoire ancienne de l'Université de Toulouse. Il faut d'abord dire que dès son origine le quartier universitaire Toulousain se trouvait plutôt dans le quartier dans lequel est située l'actuelle université Toulouse I. C'est uniquement après les événements de mai 1968 que les lettres et sciences humaines ont quitté ce quartier et se sont exilées au Mirail. L'Université a été fondé en 1229 pour éradiquer l'hérésie cathare, les premiers enseignements donnés furent des enseignements de théologie et donnés par des professeurs issus de la récente Université de Paris. L'Université au départ était à la fois un symbole de l'inféodation au dogme catholique mais aussi de l'inféodation au pouvoir de la France et de Paris. Mais très vite après sa création l'Université de Toulouse s'est orienté vers le droit et cette discipline est devenue plus importante que la théologie. Les statuts de l'Université de Toulouse se sont indexés sur ceux de l'Université de Bologne, la plus ancienne université d'Europe, créée en 1089 et lieu de «redécouverte» du droit romain. Le droit est resté jusqu'au début du vingtième siècle la discipline la plus importante, numériquement parlant. L'importance de la justice à Toulouse a été encore plus grande quand, en 1420, a été instauré le Parlement de Toulouse, le parlement c'était la cour royale de justice.

4 Droit et mathématiques, équité et égalité.

Cette digression sur le droit dans un exposé consacré à la modélisation statistique est surtout là pour insister sur les liens historiques entre calcul et justice. Une expression telle que «deux poids, deux mesures» montre que ce lien existe encore dans le langage français courant. Les poids et mesures évoquent le monde la physique et du calcul, alors que l'expression se réfère à une injustice. On dit aussi une lourde peine ou une lourde amende, comme si la justice pesait les fautes et infligeait des peines en fonction de ce poids, c'est sans doute pour cela que la balance reste le symbole de la justice alors que c'est surtout un outil de scientifique, physicien, chimiste, biologiste etc. Le mot mesure, comme beaucoup de mots, a plusieurs sens, pour les statisticiens de formation mathématique il évoque la théorie de la mesure qui est un préalable incontournable pour faire des probabilités qui sont elles mêmes incontournables pour faire des statistiques. Attention, toutefois, les statistiques élémentaires étaient pratiquées à Rome, durant l'antiquité Chinoise et même plus anciennement, les probabilités sont apparues à la Renaissance, la théorie de la mesure a été élaborée au vingtième siècle, la chronologie du système éducatif n'est pas la même que celle de l'histoire des sciences. Il est intéressant de noter que le mot même de probabilité a une étymologie liée au droit; probable est en quelque sorte un synonyme de prouvable, ce que nous appelons maintenant calcul des probabilités s'appelait encore au dix septième siècle «géométrie des chances». Il faut bien noter que souvent c'était les juristes qui avaient le plus besoin de connaître l'utilisation des chiffres et de la géométrie utilitaire, l'économie n'est apparue que tardivement, comme une discipline indépendante du droit. Souvent parmi les «mathématiciens» de la Renaissance on trouve des spécialistes du droit ayant reçu initialement une formation de juriste. Un exemple typique est François Viète né en 1540 à Fontenay le Comte, ayant fait des études de droit à Poitiers, obtenant une charge d'avocat en 1560. Viète est surtout connu pour avoir introduit les notations mathématiques usuelles, $+$, $-$,$=$ pour l'algèbre, et l'usage de lettres pour désigner les quantités inconnues, mais il a écrit plusieurs ouvrages de mathématiques. Viète avait utilisé les premières lettres de l'alphabet, a, b, c, pour désigner les quantités inconnues dans les équations, par la suite Descartes a décidé d'utiliser les dernières lettres x, y, z et depuis l'usage est resté. Viète était conseiller parlementaire, un de ses titres de gloire est d'avoir su décoder les messages secrets espagnols en 1590, pour le compte d'Henri IV au cours d'une guerre. Nombreux sont les juristes du seizième siècle qui étaient des érudits, un des facteurs évidents de cette érudition se situant dans les progrès de l'imprimerie et la disponibilité des livres, tous n'étaient pas nécessairement capables d'innover comme le fit par exemple Viète, mais beaucoup étaient capables de comprendre les innovations des autres et même de se passionner pour celles ci. Le parlement de Toulouse au seizième siècle était un lieu d'érudition mais il fallut attendre le dix septième pour qu'un juriste conseiller au parlement de Toulouse laisse un nom dans le champ des mathématiques, mais on peut dire que cela valait la peine d'attendre, étant donné qu'il s'agit du plus illustre des mathématiciens ayant vécu dans cette ville et sans doute un des plus illustres de tous les temps.

5 Droit et mathématiques, un exemple local.

Lorsque l'on s'intéresse à l'histoire des sciences à Toulouse il est impossible de faire l'impasse sur Pierre de Fermat. En fait Fermat n'est pas né à Toulouse, mais à Beaumont de Lomagne le 17 Août 1601 et il est mort à Castres le 12 Janvier 1665, et il a fait ses études de droit à Toulouse, Bordeaux puis Orléans. C'est à Bordeaux qu'il a rencontré Jean Beaugrand, lequel avait appris les mathématiques avec Viète, c'est là que Fermat s'est d'abord intéressé aux mathématiques et a écrit les premières choses sur le sujet mais il est vrai que la quasi totalité de son oeuvre scientifique a été faite à Toulouse où il était conseiller au parlement de 1631 à sa mort. De plus Fermat avait des attaches familiales importantes à Toulouse, deux des frères de sa mère ayant été capitouls. Une vision classique tend souvent à présenter les mathématiciens du dix septième siècle comme des nobles désoeuvrés résolvant des problèmes comme s'il s'agissait de jeux. Cette vision ludique est parfois renforcée pour les problèmes de probabilité qui étaient souvent liés à des jeux de hasard en vogue à l'époque... Fermat n'était pas d'origine noble, il est né Pierre Fermat, il s'est anobli en s'enrichissant en étant parlementaire, ce qui était un schéma classique depuis le Moyen-Âge à Toulouse mais surtout dans les villes italiennes. En ce qui nous concerne dans cet exposé Fermat a été connu par sa correspondance avec Blaise Pascal (1623-1662) et notamment la lettre de 1654 concernant la géométrie des chances c'est à dire les probabilités. Il a été surtout connu par l'ensemble des mathématiciens pour son fameux théorème qui n'a été démontré que le 23 Juin 1993, lors d'un séminaire à Cambridge par un mathématicien Anglais, Andrew Wiles. En fait il y avait un trou, dans la démonstration et le théorème n'a été considéré comme démontré qu'en 1994, jamais un théorème n'a fait autant de bruit. Fermat, contrairement, à ce que l'on dit habituellement avaient des contacts avec des mathématiciens à Toulouse, son collègue au parlement Pierre de Carcavi s'intéressait beaucoup aux mathématiques et c'est en grande partie grâce à Carcavi que Fermat a pu rentrer en contact avec les savants de son temps, Pascal, Roberval, Huygens. Carcavi était originaire de Lyon et après 1636 il alla à Paris où il rencontra Mersenne, un moine mariste qui rêvait de mettre en contact tous les scientifiques du monde (alors qu'il n'y avait pas encore Internet...). Fermat a aussi fait des mathématiques, avec le père Antoine Laloubère qui enseignait cette discipline, en même temps que le grec et l'hébreu dans le collège des jésuites de Toulouse. Ce collège a été créé en 1567 et à l'époque de Laloubère et Fermat il y avait plus de 1200 élèves. Les travaux publiés par Fermat et Laloubère portaient sur la géométrie, la cycloïde de Laloubère. Il est intéressant de noter que les locaux dans lequel se trouvait le collège des jésuites, abrite maintenant un lycée qui s'appelait, jusqu'au 22 Juin 1957, Lycée National de Garçons de Toulouse mais qui à partir de cette date s'est appelé (et s'appelle toujours) Lycée Pierre de Fermat. Il est donc vraisemblable que Fermat soit allé un jour faire de la géométrie avec un jésuite dans un lieu qui porte maintenant son nom. Fermat a eu aussi de célèbres polémiques avec Descartes, cela est même mentionné dans la toute nouvelle fresque sous les arcades du Capitole, à côté du portrait de Fermat. Anecdotiquement le portrait et la plaque indiquant quelques informations se trouvent exactement devant le café Florida, ce qui nous renvoie à l'exposé du 16 Janvier 1998 consacré à des données sur les peines de mort prononcées dans l'état américain qui porte ce nom, en version originale. Il y a eu un conflit de paternité entre les deux savants sur l'idée de mettre en connexion, la géométrie et l'analyse, on appelle cela une représentation Cartésienne. Lors d'un échange de correspondance entre les deux hommes, Descartes avait refusé de répondre à Fermat et avait dit «Monsieur Fermat est un Gascon, moi pas!». Cette phrase est resté célèbre et symbolise souvent pour les Gascons et la Gascogne, le mépris qu'ont à leur égard les gens de là haut... Descartes avait sans doute une préférence pour le nord, il vécut vingt ans en Hollande et finit sa vie à Stockholm. Mais je le répéte Fermat nous intéresse aujourd'hui en tant que géomètre des chances ou probabiliste.

6 Les probabilités, une place à part dans les sciences.

Les probabilités sont essentielles dans la modélisation statistique des données et pourtant il faut reconnaître que les probabilités occupent une place particulière dans le champ des mathématiques. «Le calcul des probabilités ne me semble avoir été réellement, pour ses illustres inventeurs, qu'un prétexte commode à d'ingénieux et difficiles problèmes numériques... Quant à la conception philosophique sur laquelle repose une telle doctrine, je la crois radicalement fausse et susceptible de conduire aux plus absurdes conséquences... C'est la notion fondamentale de la probabilité évaluée, qui me semble directement irrationnelle et même sophistiquée: je la regarde comme impropre à régler notre conduite en aucun cas, si ce n'est tout au plus dans les jeux de hasard... Les applications utiles qui semblent lui être dues, le simple bon sens, dont cette doctrine a souvent faussé les aperçus, les avait toujours clairement indiquées d'avance.» (Auguste Comte. Cours de philosophie positive, 27ème leçon 1844). «Statistique n'est pas probabilité. Sous le nom de statistique mathématique des auteurs, qui, je vous le dis en français n'écrivent guère dans notre langue, ont édifié une pompeuse discipline riche en hypothèses qui ne sont jamais satisfaites.» (Jean-Paul Benzécri, l'Analyse Des Données, 1972). Benzécri est l'inventeur de la méthode qui s'appelle l'analyse factorielle des correspondances (AFC) et le fondateur de l'école française d'analyse des données Ces deux citations montrent le scepticisme de certains auteurs, philosophes ou mathématiciens à l'égard de l'utilisation des probabilités. Le groupe de mathématiciens nommé Bourbaki, très important dans la redéfinition des mathématiques au vingtième siècle a superbement ignoré pendant très longtemps cette discipline. Les probabilités ne font pas non plus partie du cursus des élèves en classes préparatoires scientifiques, elles ne sont enseignées que pour les biologistes, vétérinaires, économistes, mais pas pour les élèves se dirigeant vers Polytechnique, les Mines, Centrale ou Normale Supérieure. Laurent Schwartz, dans son autobiographie «Un mathématicien aux prises avec le siècle» (Odile Jacob 1997) parle de ce phénomène d'exclusion des probabilités, il se trouve que son frère, Daniel Schwartz, est un des statisticiens les plus connus, un des spécialistes mondiaux de l'épidémiologie, et que Laurent Schwartz est le gendre de Paul Lévy qui fut un des probabilistes les plus importants de ce siècle.

7 Fragments d'histoire des sciences, le hasard.

Il semble qu'une des raisons simples pouvant expliquer cette mise à l'écart des probabilités réside dans l'importance des mathématiques et des mathématiciens grecs sur la mathématique contemporaine. Les grecs n'ont jamais essayé de mêler les mathématiques et le calcul sur les chances, alors qu'il y a eu de très brillants géomètres dans la civilisation grecque il n'y a jamais eu de géomètres des chances, les grecs n'ont jamais vraiment essayé de «comprendre» ce que nous appelons le hasard. Le mot hasard vient de l'arabe az-zart qui veut dire le dé, les phénomènes dus au hasard s'appellent les aléas ou phénomènes aléatoires, le mot aléa vient du latin alea qui veut dire les dés. Le jet de dés est typiquement le symbole du hasard, ce jeu est pratiqué depuis la haute antiquité, comme le jet de pièce de monnaie. Les grecs et les romains jouaient abondamment aux dés. La pratique de jeux de hasard avait souvent des enjeux d'argent, mais c'était aussi très souvent lié à des pratiques divinatoires, comme le Yi-King chez les taoïstes chinois, mais aussi chez les grecs, les égyptiens etc... La volonté de connaître l'avenir a toujours hanté les civilisations, mais la scientifisation des prédictions ou plutôt la volonté, la tentative de scientifisation des prédictions n'a pas mené dans la mathématique grecque aux prémisses du calcul des probabilités. Pourtant on retrouve chez les romains en 284 après JC, avec Dioclétien des tentatives de budget prévisionnel, bien éloignés des modèles économétriques de notre époque. En effet les mots modèle et modélisation ont très souvent une connotation économique. À la fin du Moyen-Âge Européen, les jeux de hasard prennent une plus grande importance que par le passé et sans doute une signification différente. C'est en 1377 que sont imprimées les premières cartes à jouer en Europe, les cartes à jouer sont comme bien d'autres objets usuels une invention bien plus ancienne des chinois, le papier, l'imprimerie, la boussole, la poudre à canon etc.. Les cartes à jouer permettent de diversifier et de complexifier grandement les jeux de hasard.

8 Préhistoire des probabilités.

D'un autre côté apparaît en 1494 un livre extrêmement important pour les mathématiques de l'époque «Summa de arithmetica, geometria proportioni et propornionalita» écrit par Luca Pacioli (1445-1517). En fait ce livre est une compilation des mathématiques connues en ce temps là, il contient des éléments de géométrie hérités des grecs, l'algèbre et l'arithmétique héritées des arabes et des hindous et intégrées dans le Moyen-Âge européen par notamment Léonard de Pise Fibonacci (1170-1250). Mais surtout ce livre contient des éléments de comptabilité (le système de comptabilité à double entrée) de techniques de calculs, et la première version connue du problème des parties qui ne sera considéré comme résolu qu'en 1654 lors de l'échange de courrier entre Pascal et Fermat. Dans une optique scientifique chauvine française c'est souvent la date de 1654 qui est donnée comme date de naissance du calcul des probabilités. Cette vision française ignore le travail des précurseurs italiens, notamment Jérôme Cardan (Girolamo Cardano) (1501-1576). Vers 1564 Cardan a écrit «De ludo aleae» un livre qui n'a été connu que bien plus tard, mais consacré comme son nom l'indique aux jeux de hasard, il existe une formule élémentaire de probabilité que l'on appelle formule de Cardan. Si les travaux de Cardan ont parfois ont été ignorés par contre il s'est vu attribuer des inventions qu'il a piratées, le système de suspension qui s'appelle le cardan en honneur de Cardan a été inventé par les chinois et était connu en Europe vers le treizième siècle. L'attribution des découvertes scientifiques ou techniques est très souvent fantaisiste, on voit parfois attribuer à Pascal l'invention de la brouette, elle aussi connue par les chinois quelques siècles avant Pascal... Outre Cardan un certain nombre de mathématiciens italiens de la Renaissance se sont intéressés aux probabilités avant que des français s'y intéressent, on peut citer Tartaglia mais aussi Galilée qui a résolu le fameux problème dit du «Grand Duc de Toscane». L'étude des mathématiques de la Renaissance Italienne permet de voir le lien, mentionné plus haut, entre droit et mathématiques. Bologne qui était une ville, pour qui le droit avait été essentiel dans le démarrage de son Université a été un haut lieu au seizième siècle des mathématiques, Pacioli, Cardan ont été professeurs à Bologne c'est dans cette ville qu'a été officiellement trouvé la solution générale des équations du troisième degré ainsi que celles du quatrième degré. À partir de telles réflexions on peut voir aussi que le terme Renaissance n'est peut être pas adaptée car les sciences de cette époque se sont largement détachées de la simple redécouverte du passé gréco-romain, l'intégration des savoirs du Moyen-Âge et des savoirs des civilisations musulmane, hindoue ou chinoise a joué un rôle important.

9 Une énigme de l'histoire des sciences.

La question que l'on peut, légitimement se poser est : pourquoi les mathématiciens grecs de l'antiquité n'ont pas entrepris la rationalisation du hasard, alors que leur savoir mathématique était grand et nous sert souvent de référence et que les calculs sur les chances sont parfois particulièrement simples ? Pourquoi a-t-il fallu attendre la fin du Moyen-Âge européen pour voir une telle entreprise ? Premièrement cette question est similaire à la question: «pourquoi les mathématiciens grecs n'ont ils jamais utilisé le zéro ?». On peut avancer l'idée que le système de numération et les chiffres indiens (appelés chiffres arabes) a grandement facilité les calculs, mais il est vrai que le zéro a une utilité réelle en probabilité, par exemple on dit qu'un événement impossible a une probabilité nulle c'est à dire égale à zéro. Attention, toutefois la réciproque n'est pas vraie, un événement de probabilité nulle n'est pas nécessairement impossible... Il semble aussi que le concept d'infini ait été nécessaire aux premières élaborations de l'idée de probabilité. Je vais essayer de donner une définition naïve, intuitive du concept de probabilité, attention une telle définition est fausse au yeux de la théorie moderne des probabilités. Supposons que l'on jette un dé et que l'on s'intéresse à l'apparition de la face «six», celle qui est marquée du chiffre 6 ou plutôt de six points. Si l'on jette le dé un grand nombre de fois on peut s'intéresser au nombre de fois où cette face est apparue sur le haut du dé et comparer ce nombre au nombre de fois où le dé a été jeté. Si on fait le rapport entre les deux nombres on obtient la proportion de fois où le six est apparu. En principe on doit avoir la perception que lorsque le nombre de jets du dé devient très grand la proportion se stabilise autour d'un nombre. Autrement dit lorsque le nombre de jets du dé tend vers l'infini, la proportion du nombre de fois où c'est le six qui est apparu tend vers un nombre qui représente la chance d'apparition du six et que l'on appelle probabilité d'apparition du six. Dans cette vision naïve le concept de probabilité d'un événement est lié à l'idée d'une répétition infinie. L'infini n'a pas été utile uniquement aux probabilités, il a été fondamental pour le développement de l'analyse, calcul différentiel, calcul intégral, on appelle parfois cela le calcul infinitésimal, mais aussi pour la géométrie. Le concept d'infini a aussi joué un rôle fondamental dans l'art, il est nécessaire d'avoir une idée de l'infini pour faire de la perspective en peinture, le point de fuite dans un tableau avec une vue perspective est une sorte de point à l'infini. Les premiers concepteurs de tableaux avec des perspectives étaient souvent des techniciens, des ingénieurs et mathématiciens, le prototype de ces artistes scientifiques est Léonard de Vinci (1452-1519) qui a beaucoup travaillé avec Luca Pacioli ainsi qu'avec le père de Jérôme Cardan. En dehors de Léonard il faut citer Léon Battista Alberti (1404-1472) et Piero della Francesca (1412-1492), Tous les deux ont travaillé sur la géométrie et fait de la peinture et surtout mis au point le concept de perspective en peinture. Les interconnexions entre savants et artistes de cette époque sont très nombreuses. Même au vingtième siècle il y a parfois des liens entre art et sciences, le peintre Escher en est un bon exemple, Dali, peintre de génie, a aussi émis des idées sur le positionnement du centre du monde. Ce n'est pas seulement dans les sciences et dans l'art que l'idée de l'infini a joué un rôle, elle a semble-t-il aussi aidé à modifier la représentation du monde chez certains philosophes, comme Giordano Bruno par exemple. Mal lui en a pris, sa vision infinie de l'univers n'a pas eu l'heur de plaire à l'église et en 1600 il a été brûlé vif à Rome comme hérétique, avant cette funeste aventure il avait enseigné brièvement la théologie à l'Université de Toulouse. Il ne parait pas déraisonnable, toutefois de considérer qu'une vision infinie du monde permette plus facilement de se détacher du modèle géocentrique, le rejet du modèle géocentrique est ce qui a valu un procès resté célèbre avec Galilée comme accusé (pourtant elle tourne). En dehors du zéro, des chiffres et de l'infini, il y a une autre absence mathématique importante chez les mathématiciens grecs de l'antiquité c'est ce que j'appellerai la mathématique du temps et du mouvement. Pour expliquer cela il suffit de rappeler le fameux paradoxe d'Achille et la tortue. Ce paradoxe a été imaginé par Zénon d'Élée (490-430 Avant JC). Achille fait une course avec la tortue. La tortue a 100 mètres d'avance. Maintenant, dit Zénon, Achille parcourt 100 mètres et atteint le point de départ de la tortue. Pendant ce temps la tortue a fait le dixième du chemin parcouru par Achille, et se trouve maintenant à 10 mètres devant Achille. Achille parcourt ces 10 mètres. Pendant ce temps la tortue a parcouru 1 mètre. Achille parcourt ce mètre; la tortue avance de 10 centimètres. Achille, parcourt ces 10 centimètres; la tortue avance d'un centimètre. Achille parcourt ce centimètre; la tortue avance d'un millimètre. Ainsi arguait Zénon, Achille se rapproche constamment de la tortue mais il ne peut jamais la rattraper. La présentation de ce paradoxe hyper connu est tiré de «Les mathématiques pour tous» de Lancelot Hogben, bibliothèque scientifique PAYOT 1962, un excellent livre de vulgarisation mathématique, qui est un livre extrêmement connu dans le monde anglophone sous son titre original «Mathematics for the million» datant des années 1930. Lancelot Hogben était professeur à Aberdeen en Écosse. Bien évidemment les grecs savaient qu'Achille ne pouvait que dépasser la tortue mais ils étaient incapables de poser le problème correctement, au sens où nous l'entendons actuellement. La géométrie grecque paraît de nos jours étrangement statique, le temps n'apparaît pas comme une variable importante dans la position des problèmes de mathématiques et surtout le temps n'y est pas représenté de manière continue comme dans toute l'analyse moderne depuis la Renaissance.

Or il est évident que l'existence même du calcul des probabilités est lié à la prédiction et donc au temps futur, la compréhension des premiers calculs des probabilités a été contemporaine de la compréhension moderne du temps en mathématique. Toutes ces raisons avancées peuvent se résumer en une seule, l'absence des outils mathématiques nécessaires aux probabilités ont empêché les mathématiciens grecs de l'antiquité d'aborder un tel sujet.

10 Un autre type d'explication.

D'un autre côté on peut se demander si la vision du hasard dans l'antiquité gréco-romaine n'était pas différente de celle qui a vu le jour vers la fin du Moyen-Âge européen et du début de la Renaissance pour des raisons de différences de «mentalité». Dans l'antiquité il y avait plusieurs dieux mais même Zeus le dieu le plus important des grecs était soumis à une instance supérieure qui s'appelait le Destin, à fortiori les mortels se considéraient comme prisonniers du Destin. La divination envisageait, semble-t-il, de connaître son avenir mais pas vraiment de le changer. À l'intérieur de la mythologie, il existe un personnage dont l'histoire caractérise cet aspect inéluctable du Destin, il s'agit d'Oedipe. Laïos, roi de Thèbes et Jocaste se firent annoncer par la Pythie de Delphes que leur fils Oedipe tuerait son père Laïos et épouserait sa mère Jocaste. À cause de cet oracle les parents se séparèrent d'Oedipe qui fut récupéré et élevé par le roi de Corinthe. Devenu grand Oedipe rencontra un devin, aveugle mais clairvoyant, qui lui refit la terrible prédiction, qu'il ignorait, il fuit ceux qu'il croyait être ses parents. Sur la route de Thèbes il se dispute avec un étranger et le tue, il s'agit de Laïos son père. Plus tard il rencontre le Sphinx et résout l'énigme du Sphinx, ce qui lui vaut de devenir roi de Thèbes et de gagner la main de Jocaste. L'oracle de la Pythie s'est donc réalisé malgré les efforts des mortels, le Destin est inéluctable. Le mythe d'Oedipe a été utilisé par la psychanalyse mais pas vraiment pour marquer le caractère non négociable de la destinée humaine, mais plutôt, dans ma vision naïve de la psychanalyse, pour parler des relations entre parents et enfants. Le caractère inéluctable et non négociable du futur, peut être, c'est une hypothèse personnelle, une raison pour ne pas chercher à le rationaliser. Un certain nombres d'ouvrages sur le Moyen-Âge montrent qu'une hypothèse de ce style peut toutefois être envisagée. Beaucoup plus tard Einstein qui refusait l'utilisation des probabilités dans la mécanique quantique avait dit «Dieu ne joue pas aux dés»; une telle attitude au vingtième siècle de la part d'un aussi grand savant montre qu'une telle hypothèse peut avoir un peu de sens. Les convictions religieuses, idéologiques où politiques peuvent, semble-t-il, influencer le travail des scientifiques qui sont après tout, aussi, des êtres humains soumis à des croyances.

11 Exemple de raisonnement du type modélisation.

Ce que je viens de présenter est, si on peut dire conforme, à ce que l'on pourrait appeler l'esprit de la modélisation statistique appliqué à un problème d'histoire des sciences. La variable à expliquer est l'apparition du calcul des probabilités, ou plutôt le contraire c'est à dire, la non apparition du calcul des probabilités dans l'antiquité grecque. Dans un premier temps on s'est posé la question de savoir si cette non apparition pouvait être expliquée par la non existence des outils mathématiques adéquats. Ce premier modèle interprétatif peut être qualifié d'internaliste, l'histoire des sciences étant considérée comme un champ autonome. Dans un deuxième temps, nous nous sommes demandés si la représentation de la destinée humaine par la civilisation grecque antique pouvait influencer la représentation du hasard et, en quelque sorte, interdire la mathématisation de celui ci. Ce deuxième modèle interprétatif peut être qualifié d'externaliste, l'histoire des sciences y est considérée comme dépendante des autres activités humaines. En fait les deux modèles ne se contredisent pas nécessairement, ils sont, sans aucun doute, «faux» tous les deux ou pour le moins largement incomplets et il y a sans aucun doute bien d'autres modèles explicatifs. On peut dire que c'est d'ailleurs l'intérêt essentiel de cette démarche que de pouvoir essayer d'autres modèles. J'ai volontairement ignoré le modèle économiste «marxiste» traditionnel qui dirait grossièrement que les conditions économiques de l'antiquité n'étaient pas favorables aux mathématiques de la spéculation économique. À contrario la fin du Moyen-Âge même dès le douzième siècle, la vie urbaine devient plus importante et ce sont les banquiers et les marchands qui y détiennent le pouvoir, cela est par exemple visible à Toulouse avec l'apparition des Capitouls, mais c'est encore plus net dans les républiques Italiennes, Gênes, Pise, Florence etc. À Florence l'ascension des Médicis marchands et banquiers est symbolique de ce phénomène. Il existe une importante littérature sur l'histoire économique et sociale de la fin du Moyen-Âge, des points d'entrée faciles se trouvent dans bien des livres de Jacques Le Goff. Il est vrai que les pratiques commerciales, économiques et bancaires de la fin du Moyen-Âge apparaissent différentes de celles de l'antiquité qui étaient basées sur l'esclavage et utilisaient très peu la banque, l'assurance et la spéculation financière. Attention, quand même, on trouve des traces de pratiques économiques modernes dans l'antiquité. Le modèle économiste fournit un autre type d'explication, en terminologie statistique, l'économie apparaît comme la variable explicative la variable à expliquer étant toujours l'émergence du calcul des probabilités. Si l'on considère que le développement du calcul des probabilités est une des multiples composantes du développement scientifique de la Renaissance il est évident que des tas d'autres modèles peuvent être envisagés. Dans un livre intitulé «The emergence of probability» Ian Hacking Cambridge University Press (1975), l'auteur recense cinq explications possibles. J'ai trouvé un commentaire sur la question dans le livre «L'empire mathématique» de Phillip J. Davis et Reuben Hersh chez Gauthier Villars (1988) dans un passage intitulé «l'univers stochastisé» Les deux auteurs soulèvent le problème que j'évoque en une phrase que je cite: «Le surgissement tardif de la théorie des probabilités est une des énigmes de l'histoire des sciences.». Ils ajoutent qu'il y a une coïncidence entre l'apparition de la mécaniste déterministe et d'une vision probabilisée du monde, deux choses apparemment contradictoires... En plus de «L'empire mathématique» Phillip J. Davis et Reuben Hersh ont écrit «L'univers mathématique» chez Gauthier Villars (1985), les deux ouvrages ont été traduits de l'anglais (plutôt de l'américain) et les versions originales sont plus anciennes, les deux livres sont d'excellents ouvrages de vulgarisation mathématique et informatique, mais les parties concernant l'informatique sont un peu datées, celles consacrées aux mathématiques ont mieux vieilli. Il est par exemple très intéressant pour un historien géographe de regarder le rôle particulier qu'a joué Venise au Moyen-Âge et au début de la Renaissance. Le livre de Luca Pacioli summa... a été imprimé en 1494 à Venise comme près de la moitié des livres imprimés à cette époque. Les pratiques financières de la république de Venise étaient en avance sur leur temps et certaines apparaissent dès le neuvième siècle. Le problème des parties, déjà mentionné, peut être considéré comme Vénitien car il est connu avec une unité monétaire qui est le ducat, c'est à dire la monnaie de Venise. Les Vénitiens à cause de leur possessions maritimes étaient en contact avec les civilisations orientales et tout le monde a entendu parler de Marco Polo célèbre voyageur Vénitien. La prise de Constantinople par les Turcs en 1452, c'est en fait symboliquement la fin du Moyen-Âge, a ramené énormément de savants grecs vers Venise et en raison de sa puissance industrielle, liée à ses gigantesques arsenaux, disposait, au quinzième siècle, d'une forte présence d'ouvriers, de techniciens et d'ingénieurs. Les raisons matérielles ont favorisé l'explosion de l'imprimerie, la présence d'«intellectuels» a permis de faire des livres... Venise a ainsi acquis un rôle essentiel dans l'histoire des sciences à la fin du quinzième siècle, époque de l'apparition du calcul des probabilités ou plutôt des signes avant coureur des probabilités. Si on regarde les époques antécédentes on trouvera que l'Andalousie ou la Catalogne ont eu au moment des grandes traductions, au douzième siècle, un rôle essentiel tant dans la redécouverte des mathématiques grecques que dans l'assimilation des mathématiques arabes. Les Juifs d'Espagne ont joué un rôle très important, dans ce processus et leur expulsion d'Espagne en 1492 a permis de diffuser leur savoir à travers l'Europe. Il est inutile d'essayer de trouver tous les modèles possibles et toutes les sources possibles d'explications à un phénomène, mais il est sûr que tenter d'en trouver plusieurs, est à mon humble avis préférable à se fixer sur un seul modèle. Attention à tous les modèles que je viens de mentionner il ne sont que des réflexions d'un amateur, il y a des spécialistes sérieux sur tout un tas de points que j'ai abordés, y compris dans l'Université de Toulouse le Mirail, j'en ai rencontrés. Il faut aussi considérer que la question posée est volontairement restreinte à l'antiquité gréco-latine, il y a d'autres formulations de la question qui auraient pu amener des hypothèses différentes. Une question intéressante serait par exemple : «Pourquoi la mathématique chinoise n'a pas mis au point le calcul des probabilités ?»

12 Mais qu'est ce qu'un «bon» modèle ?

Attention en statistique, ce n'est parce qu'un modèle est bon au sens des critères usuels des statisticiens, dont j'ai parlé dans le séminaire précédent, que le modèle a une quelconque validité, l'interprétation d'un modèle statistique doit toujours se faire en regardant le sens des variables utilisées. Une phrase telle que «X explique Y» , ou «la variable indépendante explique bien la variable dépendante» sont vides de sens, la deuxième est même absurde d'un point de vue réthorique, une variable indépendante qui explique c'est un oxymoron comme sombre clarté ou silence assourdissant... Il est très important, à mon avis, de savoir nommer correctement les variables, de manière à savoir ce qu'elles représentent sur le plan concret, il est, de même, important de savoir nommer les modalités pour les variables qualitatives. Dans la terminologie du logiciel SPSS, qui a d'ailleurs été reprise par d'autres logiciels, mettre des étiquettes sur les variables et les modalités s'appelle faire un dictionnaire. Faire le dictionnaire, est à mon avis, une étape essentielle, dans un traitement de données en sciences humaines (ou autres). Un tel sujet nécessiterait un développement important mais nous éloignerait trop des objectifs de cet exposé. Je conseille encore la lecture de «La mal-mesure de l'homme» de Stephen Jay Gould, Odile Jacob, 1997, dans lequel sont mentionnés quelques exemples de modèles qui marchent d'un point de vue statistique mais qui n'ont aucun sens concret. Stephen Jay Gould cite par exemple le fait que l'on puisse expliquer le prix de l'essence par son âge, quand son âge augmente le prix de l'essence augmente et la corrélation est forte, de même le prix du beurre peut être expliqué par la distance entre les étoiles de notre galaxie, plus les étoiles s'éloignent les unes des autres et plus le beurre est cher, c'est vrai mais cela ne veut rien dire même pour un astrologue. De manière réciproque ce n'est pas parce qu'un modèle étudié est «mauvais», qu'il n'est pas pertinent, qu'il n'apporte pas d'information. Un décalage entre un modèle et les observations peut être une appréciable source d'information. C'est d'ailleurs une démarche intéressante dans la recherche, je cite pour cela deux exemples tirés du l'autobiographie de Laurent Schwartz, voici un extrait du livre: «Au cours d'une recherche, il arrive qu'un accident se produise et jette le trouble. Si l'on se contente d'en être ennuyé, ce n'est généralement pas suffisant. Il faut chercher d'où vient le trouble, ce qui ici encore, nécessite une certaine curiosité. Il arrive qu'on s'aperçoive d'un fait beaucoup plus important que toute la recherche qu'on faisait antérieurement. Il faut alors abandonner la première recherche pour passer à la seconde qui donnera des résultats bien plus fructueux. C'est une chose courante. La recherche nécessite donc un esprit toujours mobile, toujours original, toujours prêt à des révolutions, et très opiniâtre pour aller jusqu'au bout des difficultés. Elle exige le courage et la curiosité. Je donne couramment l'exemple de Fleming, ce biologiste qui étudiait des cultures bacillaires. Un jour il s'aperçut que l'une de ces cultures avait été envahie par des champignons. Au lieu de la jeter, il eut la curiosité de l'examiner, et il vit que tous les bacilles avaient disparu. Il eut le courage d'abandonner sa recherche antérieure pour y réfléchir. Ce fut la découverte de la pénicilline; la pénicilline était un produit du champignon qui tuait les bacilles, c'est une méthode de défense du champignon. La découverte de la pénicilline a été beaucoup plus importante que celle sur laquelle travaillait initialement Fleming. Tout le monde n'est pas Fleming, mais il est bon d'avoir une pareille curiosité. On peut aussi rappeler que Christophe Colomb, obsédé par l'idée de trouver la route occidentale des Indes découvrit l'Amérique.» L'idée que la terre est ronde était fort ancienne, on la doit à Pythagore (580-500 avant JC), mais l'idée d'en faire le tour fut tardive, faute de moyens suffisants; le modèle utilisé par Colomb n'était pas absurde mais comportait deux «erreurs» importantes. Une erreur sur les dimensions de la sphère terrestre et surtout, la présence non prévue de terres importantes barrant la route des Indes. En modélisation statistique des données on a souvent à faire face à des situations qui ont une certaine analogie avec la situation de Colomb. Par exemple dans une régression linéaire une valeur qui ne s'ajuste pas à la droite de régression est appelée valeur aberrante, cette appellation pose problème, il est possible que pour un chercheur, ce soit justement cette valeur qui offre le plus d'intérêt et qui soit le point de départ d'une nouvelle étude. Le phénomène fréquent qui se produit, chez les utilisateurs de la statistique, est de croire à la valeur des modèles en raison des mathématiques qui sont utilisées dans le modèle. Stephen Jay Gould le dit de la façon suivante :« L'esprit de Platon a la vie dure. Nous sommes incapables d'échapper à cette tradition philosophique qui veut veut que ce que nous voyons et mesurons dans le monde ne soit que la représentation superficielle et imparfaite d'une réalité cachée. L'essentiel de la fascination des statistiques tient dans ce sentiment viscéral - Méfiez vous toujours des sentiments viscéraux - que les mesures abstraites résumant des grands tableaux de données doivent exprimer quelque chose de plus réel et de lus fondamental que les données elles mêmes. Pour acquérir un bon niveau professionnel, les statisticiens doivent faire un effort conscient pour contrebalancer cette tendance naturelle.»

13 La revanche de Platon?

Je partage pleinement le point de vue de Stephen Jay Gould, mais je précise qu'il faut, à mon avis, se méfier du rejet complet de l'esprit de Platon, et je profite de cette remarque pour dire combien je suis en désaccord avec les propos sévères contre les mathématiques de Claude Allègre dans son livre intitulé «La défaite de Platon». La valeur des statistiques et des mathématiques en général ne réside pas dans le fait qu'elles révéleraient quelques choses de plus réel que le monde perceptible. Un modèle statistique est une entité mathématique qui résume des données, il n'a pas plus de valeur que les données elles mêmes, il est surtout plus facile à manipuler. Les attaques de Claude Allègre contre les mathématiques et contre l'utilisation de celles ci font aussi l'objet de déclarations publiques. Voilà par exemple un extrait d'un article écrit par Claude Allègre paru dans le Monde du vendredi 6 Février 1998, article ayant l'honneur de la première page et intitulé modestement «Ce que je veux».: «Naguère, on pensait que les mathématiques régnaient au-dessus des sciences. Aujourd'hui on constate que les grandes sciences (biologie, informatique, chimie par exemple) se développent en dehors d'elles. Il n'y a pas si longtemps encore, on tenait les sciences de la nature pour de pures descriptions qualitatives. Aujourd'hui ce sont les sciences de la vie ou de la Terre qui occupent le devant de la scène. Le point de vue historique s'impose en sciences de la matière, de la vie, comme de la société.» Je me refuse de faire un commentaire critique détaillé de ce paragraphe mais il correspond à une sorte d'antithèse exacte de ce qu'est mon travail. En tant que mathématicien appliqué, j'interviens auprès des spécialistes des grandes sciences pour les éclairer avec le savoir d'une science présentée comme légèrement dépassée. Il se trouve et c'est ce qui est gênant, que monsieur Allègre ne tient pas ce discours en tant que spécialiste des sciences de la Terre mais en tant que ministre de l'éducation nationale, de la recherche et de la technologie. En tant que fonctionnaire de l'éducation nationale, c'est donc mon ministre de tutelle et je suis donc tenu de ne pas trop étaler mon opinion négative. Parfois un phénomène physique ou biologique peut apparaître comme complètement indépendant de la mathématique de l'époque où il a été découvert puis les mathématiques évoluant, le phénomène peut être modélisé par les mathématiques. L'exemple auquel je pense c'est le mouvement brownien, ce phénomène a été découvert par un botaniste écossais nommé Robert Brown (1773-1858), à l'époque de Brown ce phénomène était quelque chose de purement «botanique» concernant les pollens, le mouvement brownien a été identifié en 1827. Mais l'étude du mouvement brownien est devenu un domaine très important de la théorie moderne des probabilités, notamment grâce aux travaux de Paul Lévy (1886-1971), le livre de Paul Lévy «Processus stochastiques et mouvement brownien» date de 1948. Entre 1827 et 1948 le mouvement brownien est passé de la botanique aux mathématiques, alors que sa première apparition dans l'histoire des sciences le faisait croire en dehors des mathématiques.

14 Digressions sur la place des mathématiques.

Contrairement aux idées répandues, les mathématiques ne sont pas uniquement là pour faire vivre les mathématiciens, ou pour pratiquer une horrible sélection sur de pauvres élèves que l'on oblige à apprendre des formules et des équations... Cette vision existe, en raison, sans doute, de certains abus des systèmes éducatifs. Il est sûr que les mathématiques nécessitent un effort d'abstraction, il y a, dès que l'on fait du calcul, un mécanisme de formalisation. Ce mécanisme de formalisation existe dès l'école primaire, lorsqu'on manipule des chiffres dans une table d'addition on fait un effort d'abstraction. 3 pommes plus 4 pommes égalent 7 pommes a un sens concret, tangible alors que $3+4=7$ est une écriture synthétique et abstraite, mais elle à l'avantage d'être valide si on remplace les pommes par des oranges, des doigts, des allumettes, des francs ou des euros... Faire la somme en mathématique est une chose apparemment simple. La table d'addition est la plus simple des tables mais il est évident que dans la réalité physique ou chimique la somme n'a pas toujours exactement la même signification. Ainsi, par exemple, si on ajoute un litre avec un litre, normalement cela doit faire deux litres car $1+1=2$ mais un litre d'eau que l'on ajoute à un litre de coton hydrophile ne feront pas deux litres de volume. C'est la même chose avec un litre de sel et un litre d'eau, car le sel se dilue dans l'eau et donne une eau salée de densité plus forte. De même un volume d'oxygène avec deux volumes d'hydrogène donne un volume d'eau bien inférieur, il faut cependant qu'il y ait quelque chose de plus pour provoquer la réaction chimique. Cela marche aussi mais de manière inversée pour la mayonnaise, le volume du jaune d'oeuf plus le volume d'huile doit être inférieur au volume de la somme des deux, pour que la mayonnaise soit réussie, pour qu'elle soit bonne il faut en plus mettre le jus d'un citron pressé, mais il s'agit là de goûts personnels... Dans l'art c'est souvent la même chose on s'attend à ce qu'une collaboration entre artistes donne quelque chose de plus que la somme des «talents» des diverses individualités, pour donner encore une référence à mes goûts personnels je dirais que l'exemple qui me vient à l'esprit c'est certaines chansons des Beatles, qui ne sont pas réductibles à la somme du travail de Lennon et MacCartney. Les statisticiens pour décrire ce type de phénomène ont inventé l'interaction. Ceci étant la somme usuelle telle qu'elle s'apprend à l'école primaire est très utile et constitue une étape abstractive très formatrice. Le fait que bien des phénomènes usuels n'obéissent pas à la règle usuelle d'addition mathématique simple n'empêche pas celle ci d'être utilisée dans bien des cas; de plus les règles de composition plus compliquées que l'addition sont souvent définies à partir de généralisations de l'addition, l'exemple élémentaire étant la multiplication. Ce mécanisme de formalisation mathématique est porté à un niveau plus élevé dans les classes supérieures, et il est parfois délicat de recoller les formules mathématiques avec la réalité concrète qu'elles sont censées décrire. Il y a très souvent une approximation, plus ou moins importante, entre l'objet concret que l'on veut représenter et l'entité mathématique qui le représente; le triangle ou le rectangle du mathématicien ne sont pas ceux du géographe, de l'agriculteur, de l'architecte ou du notaire. Le rectangle du mathématicien est un objet idéalisé mais qui représente, de manière assez correcte une zone rectangulaire sur le sol, le calcul de la surface à partir de la formule mathématique est utilisable et utilisée couramment pour les applications concrètes. En statistique on utilise très souvent la loi normale ou loi de Gauss, la fameuse courbe en cloche, l'utilisation de cette loi est justifiée par des considérations de type passage «passage à limite», c'est à dire que l'on fait l'hypothèse d'un nombre infini d'observation potentielles. Le concept même de probabilité d'un événement fait intervenir, de façon implicite l'infini. L'infini est une abstraction mathématique assez difficile à appréhender et même impossible à utiliser sur un ordinateur, par exemple le nombre mathématique pi, noté π, a, en mathématique, un nombre infini de décimales, on ne connaît ce nombre avec seulement 50 milliards de décimales et quelque soit la puissance des ordinateurs on utilisera toujours ce nombre avec un nombre fini de décimales. L'infini qui est manipulé tous les jours par les mathématiciens restera toujours inaccessible aux informaticiens. La courbe de la loi normale ne touche jamais les bords, cela veut dire que par exemple, quand on dit que la taille des individus d'une population suit une loi normale, on accepte implicitement qu'il est probable de rencontrer des individus avec une taille négative alors que c'est absurde. Il se trouve que cette probabilité est excessivement faible et indistinguable du zéro pour la quasi totalité des calculs faits avec un ordinateur. Mais on peut trouver légitimement gênant une telle situation; l'effort d'abstraction pour surmonter cette gêne est plus important que pour utiliser la table d'addition ou le rectangle mais il est fondamentalement du même ordre. La justification de l'utilisation de la loi normale, ou d'une autre loi, peut paraître arbitraire, ou obéir simplement à des considérations internes à la mathématique, harmonie, logique, cohérence; il se trouve que ces raisons internes, qui existent, sont très fortement liées à d'autres raisons telle que la mise en oeuvre des calculs. La mise en oeuvre des calculs et notamment l'écriture de programmes informatiques, de logiciels, est souvent nécessité par des raisons externes, la faisabilité des calculs n'est pas nécessairement une préoccupation des mathématiciens. Une loi de probabilité qui interdirait le dépassement de certaines valeurs aurait apparemment plus de sens concret, mais rendrait les calculs très compliqués et souvent inutilisables. C'est souvent une qualité d'un modèle statistique que d'être simple et donc manipulable même si cette simplicité vient en contradiction avec certaines réalités.

15 Les modèles dits déterministes et leurs limites.

Pour aborder la notion de modèle statistique utilisant des probabilités il me semble nécessaire de comprendre ce qu'est un modèle mathématique dit «déterministe»; le plus simple qui me vienne à l'esprit est le mouvement d'un solide, une bille par exemple, que l'on lance en l'air mais pas à la verticale. Il n'est absolument pas faux de décrire le mouvement du solide de la façon suivante, dans un premier temps le solide va s'élever jusqu'à une certaine hauteur, dans un deuxième temps le solide va retomber. autrement dit, le mouvement du solide se décompose en deux étapes, l'ascension et la descente. C'est un modèle juste mais qui manque de précision, ce modèle n'indique pas quel sera le lieu du point de chute du solide ni même jusqu'à quelle hauteur le solide s'élèvera. Pour un joueur de rugby qui tente un drop ou une pénalité ou pour un artilleur qui envoie des boulets de canon c'est bien la trajectoire et surtout l'endroit où tombe le solide qui a de l'importance. En parlant d'une bombe de fabrication familiale et artisanale, Boris Vian dans «La java des bombes atomiques» disait que la seule chose qui compte c'est l'endroit où est ce qu'elle tombe. L'étude des mouvements des solides a été une question des artilleurs de la Renaissance italienne qui a beaucoup intéressé les mathématiciens de cette époque et notamment Nicolo Fontana plus connu sous le nom de Tartaglia (1500-1557), le surnom de Tartaglia lui venait du fait qu'il avait des difficultés à parler, suite à un accident dans son enfance, tartaglia signifie bègue en italien. L'accident était en fait une agression des soldats français qui en 1512 pillèrent Brescia ville où se trouvait la famille du petit Nicolo, son père fut tué et lui blessé et laissé pour mort avec une épée dans la bouche. Ce fut effectivement avec les guerres d'Italie, de François 1er, notamment que la France a pillé les richesses mais aussi les savoirs des italiens, musique, théâtre, peinture, architecture mais aussi mathématiques et c'est en partie à cause de ces guerres que le savoir scientifique italien est passé en France. C'est aussi à cette époque que la syphilis est passée de l'Italie vers la France... Le modèle proposé par Tartaglia était le suivant, le mouvement se décompose en trois étapes, une étape d'ascension en ligne droite, une étape en arc de cercle et une étape de descente en ligne droite. Ce modèle décrit le mouvement plus précisément que le précédent mais il ne permet pas de calculer précisément le point de chute du solide. Il a fallu attendre Galilée pour que le mouvement soit décrit comme suivant une parabole, courbe géométrique connue depuis l'antiquité, comme section d'un cône. La parabole reste un modèle imparfait, en fait ce modèle décrit bien mieux le mouvement d'un boulet de canon ou d'une boule de pétanque que celui d'un ballon de rugby, d'un volant de badminton ou d'une fléchette. En effet c'est un modèle qui ignore les frottements de l'air. Mais c'est uniquement avec Isaac Newton (1643-1727) que l'on comprendra pourquoi ce modèle parabolique fonctionnait. C'est en effet avec la compréhension de la gravitation universelle découverte par Newton, parait il à partir de la fameuse anecdote d'une pomme tombant d'un arbre, que la mécanique des solides a été considérée comme comprise, c'est d'ailleurs pour cela que l'on appelle cette discipline la mécanique newtonienne. Les découvertes de Newton permirent de comprendre de manière synthétique bien des travaux de ses prédécesseurs.

16 Astronomie, mathématiques, probabilités et statistiques

Si il y a un domaine où les travaux de Newton furent essentiels c'est bien évidemment l'astronomie. La loi de la gravitation universelle permit de comprendre les lois que Johannes Kepler (1571-1630) avait établi pour expliquer le mouvement des planètes. Kepler avait établi ces lois à partir des observations faites par l'astronome danois Tycho Brahe (1546-1601) dont il avait été l'assistant à Prague. Ce mécanisme est symbolique de la démarche scientifique, d'abord observation (Brahe), ensuite formalisation (Kepler) enfin théorisation (Newton). Il y a une découverte très importante dans l'histoire de l'astronomie qui est un très bon exemple de «modélisation» il s'agit de la découverte de la planète Neptune, cette histoire figure très souvent dans les cours de philosophie au lycée mais j'en ai trouvé une excellente version sur le WEB à l'adresse suivante: http://www.u-bourgogne.fr/c.nitschelm/histoire.html. Vous pouvez trouver à cette page un excellent cours sur l'histoire de l'astronomie, fait par monsieur Christian Nitschelm, pour des étudiants de DEUG scientifiques à l'Université de Bourgogne à Dijon, j'en ai extrait le paragraphe sur Neptune le voici: «La découverte de la planète Uranus par William Herschel en 1781 avait été suivie par des observations nombreuses au cours du temps et par une modélisation sous forme de tables de la trajectoire suivie par cette planète à partir des lois de la mécanique céleste. Les premiers temps, la planète nouvellement découverte sembla suivre correctement la trajectoire calculée, alors que tous les astronomes supposaient que les tables étaient rigoureusement exactes. En 1821, cependant, l'astronome français Alexis Bouvard reprit les tables et les observations et constata que quelque chose d'étrange était en train de se produire. Vers 1830, les soupçons de Bouvard se confirmèrent, la planète étant en train de s'écarter inexplicablement de sa trajectoire théorique. Il fallut cependant encore attendre quelques années de plus pour que cet écart devienne intolérable au niveau de la mécanique céleste. La conclusion devint alors évidente, le mouvement d'Uranus était perturbé par une autre planète. Vers 1841, l'anglais John Couch Adams débuta un long travail d'analyse théorique qui lui pris cinq ans. Ce travail, terminé en 1846, déterminait la localisation céleste et les caractéristiques du corps perturbateur qui influençait l'orbite de la planète Uranus. Vers 1843, le français Urbain Jean Joseph Le Verrier débuta indépendamment le même travail qu'il termina plus rapidement, également en 1846. Les deux hommes arrivèrent à la même conclusion: les irrégularités de la planète Uranus étaient dues à une planète perturbatrice encore inconnue située au-delà et dont la position à une certaine date pouvait être calculée. Sa masse et son pouvoir attracteur pouvaient également être déterminés. Alors qu'Adams se heurta à la bureaucratie anglaise de l'époque et à la mauvaise volonté de l'astronome royal George Biddell Airy, lequel refusa d'effectuer des recherches avec l'aide du matériel de l'Observatoire Royal de Greenwich, Le Verrier, qui était mathématicien et absolument pas observateur, utilisa très habilement ses bonnes relations avec l'astronome allemand Johann Gottfried Galle qui utilisait à l'Observatoire de Berlin un nouveau réfracteur très performant. Utilisant les calculs de Le Verrier, Galle, après quelques recherches infructueuses menées à l'aide de l'astronome allemand Heinrich Louis d'Arrest, découvrit effectivement le 23 septembre 1846, en collaboration avec ce dernier, un astre inconnu assez proche de la localisation calculée par le français. Cet astre, qui correspondait assez bien aux spécifications théoriques de Le Verrier et d'Adams, fut appelé Neptune. La découverte de la planète Neptune apparut comme étant une éclatante confirmation de la mécanique newtonienne. Cette dernière avait en effet permis de découvrir une nouvelle planète uniquement par le calcul. Cette constatation fit d'ailleurs dire à l'astronome français François Arago que Le Verrier avait vu Neptune avec la pointe de sa plume...» L'histoire de l'astronomie a énormément de liens avec l'histoire des statistiques et des probabilités, il est évident qu'en astronomie il y a énormément d'observations et que ces observations peuvent être entachées d'erreurs, dans le cas de l'observation de la trajectoire d'Uranus, l'écart entre la trajectoire prévue et la trajectoire observée a, pendant quelques années, été considéré comme non problématique, puis cet écart est devenu intolérable. Ce genre de raisonnement est typiquement celui de la statistique mathématique. Au niveau des histoires des individus il est sur qu'il y a eu énormément de connexions. Pour commencer citons Christiaan Huygens (1629-1695) qui a écrit un ouvrage très important dans l'histoire des probabilités : «De Ratiociniis in Ludo Aleae» paru en 1657; mais Huygens est connu pour avoir mis en évidence les anneaux de Saturne, en 1655 il découvrit Titan, le plus gros satellite de Saturne, il construisit des lunettes astronomiques et fit de fréquentes observations astronomiques de divers objets célestes. Huygens a étudié le droit et les mathématiques à l'université de Leyden, il a été, très jeune, en contact avec Descartes. À partir de l'histoire de Huygens et de son travail, dans le domaine de la physique du solide et des probabilités on peut noter que certaines similitudes de concepts, de calculs et même d'appellations subsistent entre la mécanique du solide et la statistique. La moyenne, en statistique, est un concept très proche du centre de gravité en mécanique, la variance en statistique peut être considérée comme un équivalent de l'inertie. Parfois dans les techniques d'analyses factorielles, on parle de pourcentage d'inertie, alors que d'autres parlent de part de variance. Il existe, en mécanique du solide, un théorème, qui s'appelle, théorème de Huygens, qui donne une formule de décomposition de l'inertie d'un solide; cette formule rappelle étrangement la décomposition de la variance, entre variance inter et variance intra telle qu'elle se pratique dans l'analyse de variance. L'analyse de la variance a été mise au point par Ronald Fisher (1890-1962) qui avait fait des études d'astronomie avant de s'intéresser à la théorie des erreurs puis aux statistiques. Fisher a étudié la théorie des erreurs à partir d'un livre écrit par Airy, l'astronome qui empêcha Adams d'observer Neptune à Greenwich et permit à Leverrier de le doubler. Dans les techniques factorielles, l'extraction des facteurs, appelés, aussi parfois, axes principaux d'inerties, se fait exactement comme en mécanique du solide l'extraction des axes principaux d'inertie. Les analogies entre les deux secteurs peuvent être donc notés. Huygens a étudié à Leyden, grande ville universitaire Hollandaise, c'est l'équipe de statistiques de cette université qui a écrit le «module» categories du logiciel SPSS, il se trouve que c'est avec ce module que l'on peut faire dans SPSS la plupart des analyses utilisant les techniques factorielles. Le mathématicien Allemand Carl Friedrich Gauss (1777-1855) a exploré énormément de secteurs des mathématiques et de la physique, il a laissé son nom à loi gaussienne, connue aussi sous le nom de loi normale, la forme de cette loi est connue sous le nom de courbe de Gauss ou courbe en cloche, en 1809 Gauss a publié «Theoria motus corporum coelestium in sectionibus conicis Solem ambientium» , un traité en deux volumes sur le mouvement des objets célestes le premier volume est constitué de mathématiques, dans le second volume Gauss montre comment estimer et ensuite raffiner l'estimation de l'orbite des planètes, après 1817 Gauss a cessé de faire de la théorie astronomique mais il fit des observations jusqu'à l'âge de 70 ans. Il semble que ce soit à Gauss que l'on doive la première idée de la régression linéaire. Pierre Simon de Laplace (1749-1827) fut un contemporain Français de Gauss, ses deux ouvrages majeurs sont «Traité de mécanique céleste» en cinq volumes et «Théorie Analytique des Probabilités», les français ont longtemps appelé la loi de Gauss, loi de Laplace-Gauss voire loi de Laplace, alors qu'il existe une loi de probabilité différente qui s'appelle loi de Laplace. Laplace a été connu surtout par son implication politique, il a été ministre de Napoléon Ier, il a été un des responsables de la mise en place du système métrique. François Arago (1786-1853) fut astronome à l'origine de la standardisation du système métrique, il fut lui aussi ministre mais de la deuxième république, il a eu quelques élèves prestigieux Urbain Le Verrier (1811-1877) qui «découvrit» Neptune mais aussi Adolphe Quêtelet (1796-1874). Quêtelet était un mathématicien Belge qui a appris les probabilités avec Laplace et Fourier et l'astronomie avec Arago, mais il est connu pour avoir été le premier à étudier la loi normale autrement que comme loi des erreurs. Mais on peut dire que Quêtelet a surtout eu un rôle essentiel dans l'utilisation de la statistique mathématique dans les sciences humaines, il a utilisé ses savoirs mathématiques pour étudier les crimes, il a collecté et analysé des statistiques sur la mortalité les crimes et amélioré les techniques des recensements de population. Son travail a provoqué de grandes controverses. Son ouvrage «Sur l'homme et le développement de ses facultés, essai d'une physique sociale» paru en 1835, présente une conception de l'homme moyen comme possédant la valeur centrale pour les différentes mesures, en supposant que ces mesures suivent une loi normale. Quêtelet a été considéré comme un des fondateurs de l'anthropométrie, l'histoire de ce concept est le sujet central du livre de Stephen Jay Gould, «la mal-mesure de l'homme».

Mais indépendamment des rôles des individus dans l'histoire il est important de voir que un des objectifs fondamentaux tant de l'astronomie que des probabilités est commun il s'agit de la temporalité. dans les deux disciplines le temps a un rôle primordial, une des premières utilités de l'astronomie est de mettre au point des calendriers, même si la découverte des nouvelles planètes semble avoir obéi à d'autres motivations apparemment. Le positionnement précis des objets célestes a été souvent motivé par les grandes navigations, à partir de ce que l'on voit dans le ciel on peut connaître sa latitude, la détermination de la longitude a été un des problèmes scientifiques les plus importants il a été résolu par une très grande précision dans la mesure du temps.

17 La précision, source d'incertitude.

La précision des mesures, est une des raisons fondamentales de l'apparition de la théorie des erreurs. La mécanique newtonienne prévoit excessivement bien les trajectoires des solides en mouvement, mais elle a eu ses limites, certaines de ces limites ont été montrées par la théorie de la relativité, mais de manière plus simple, prévoir l'endroit très précis de la chute d'un solide est souvent très compliqué. C'est d'ailleurs à partir de considérations sur la trajectoire de projectiles qu'est justifiée l'utilisation des probabilités dans un livre dont je me suis beaucoup servi dans mes études et après «Théorie des probabilités et statistique mathématique» de V. Pugachev, Éditions MIR, MOSCOU voici ce qu'écrit Pugachev: «Un autre exemple de phénomènes aléatoires peut être fourni par la dispersion des obus; les obus ne tombent jamais au même point, même quand le tir est effectué sur un seul et même point. Il semblerait que les conditions sont identiques pour tous les tirs. Toutefois les obus effectuent des trajectoires différentes et tombent en différents points. Il est par principe impossible de prévoir à l'avance en quel point précis tombera un obus donné pour cette raison précise que nous ne pouvons pas connaître avec une précision absolue les paramètres de l'état de l'atmosphère en tous les points de la trajectoire suivie par l'obus; or ces paramètres conditionnent les forces aérodynamiques et leurs moments qui agissent sur l'obus.» Dans «The emergence of probability», Ian Hacking cite comme raison, ayant empêché l'apparition des probabilités dans l'antiquité, le fait que les dés étaient faits à partir d'os d'animaux et manquaient grandement de régularité. Ce serait donc l'apparition de dés, fait avec beaucoup de précision qui aurait permis de faire des calculs précis sur les fréquences d'apparition et rendu possible la création des probabilités. Cette explication me parait secondaire, mais pas inintéressante, elle va, en tout cas de le sens de la liaison entre précision et probabilités. Au contraire cela montre que la contradiction soulevée, par Hersh et Davis, sur la coïncidence entre l'apparition de la mécaniste déterministe et d'une vision probabilisée du monde n'en est peut être pas une.

18 Raisonnement probabiliste en statistiques.

Afin d'illustrer ce qu'est le raisonnement probabiliste en statistique appliquée, j'emprunte un excellent exemple tiré du livre intitulé: «Éléments de modélisation et traitement primaire des données» de Sergheï Aïvazian, Lev Méchalkine et Igor Enukov (1986) éditions MIR, Moscou. Je me suis d'ailleurs inspiré de ce livre pour donner un titre à ma conférence, énormément de passages de ce livre seraient accessibles d'après moi à des étudiants de sciences humaines mais il y a des passages trop mathématiques, j'ai repris les pages 41-43 en éliminant ce qui était trop mathématique.

Alexis et Boris sont assis devant une table avec un jeu de dés, et aussi une bouteille de vodka, ils ont quatre dés. Alexis lance les quatre dés, si il obtient au moins un six, Boris doit lui donner un rouble, si par contre il n'obtient aucun six c'est lui qui doit donner un rouble à Boris. Les dés utilisés par Boris et Alexis sont des cubes à six faces, chacune portant un chiffre dessus de 1 à 6, et non de ces dés modernes utilisés dans les jeux de rôles, tel que «donjons et dragons» qui peuvent avoir 4 ou 12 ou 20 faces.

Arrive Sergueï, il observe Boris et Alexis en train de jouer, il a envie de gagner quelques roubles et il se demande si il doit prendre la place d'Alexis et miser sur l'apparition du six ou celle de Boris et miser sur la non apparition du six. Sergueï est patient il attend que Boris et Alexis aient joué 100 fois et il note qui a gagné. Il se trouve que 52 fois c'est Boris qui a gagné et 48 fois Alexis. C'est à dire que sur 100 lancers de quatre dés, il y a 52 lancers où cours desquels le 6 n'est pas apparu et 48 lancers où cours desquels le 6 est apparu. La non apparition du 6 a été plus fréquente que l'apparition du 6 et Segueï conclut qu'il doit prendre la place de Boris et miser sur la non apparition du 6. Sergueï applique un raisonnement statistique. La fréquence observée de la non apparition du 6 est $0,52$ et $0,52 > 0,50$.

Survient Piotr, lui aussi a envie de jouer et de gagner quelques roubles. Ce que remarque Piotr tout de suite c'est que les dés sont ordinaires, ont une forme géométrique régulière, et qu'ils sont symétriques et il en déduit que chacune des faces d'un dé a la même chance d'apparaître que les autres faces du dé et que donc chaque face du dé a une chance sur six d'apparaître, à partir de là en appliquant un calcul de probabilités assez simple il essaye de connaître qu'elle est la chance que le 6 n'apparaisse pas en lançant 4 dés à la fois. La probabilité que le 6 n'apparaisse pas en un lançant un dé est égale à la somme des probabilités d'apparition du 1, du 2, du 3, du 4 et du 5, soit 5/6 (cinq sixièmes). Les quatres lancers de dés étant considérés comme indépendants, la probabilité que le 6 n'apparaisse pas en lançant quatre dés est le produit des probabilités de la non apparition du 6 à chacun des dés soit $(5/6).(5/6).(5/6).(5/6) = 625/1296 = 0,482$. La probabilité d'apparition du 6 est donc $1 - 0,482 = 0,518$ et comme $0,482 < 0,518$, Piotr en déduit qu'il doit prendre la place d'Alexis et miser sur l'apparition du 6. Le raisonnement utilisé par Piotr est un raisonnement de type probabiliste.

Les deux raisonnements de Piotr et de Sergueï arrivent à des conclusions différentes, cela est heureux car cela leur permettra de jouer, mais il semble nécessaire de lever l'ambiguïté. C'est le but de la statistique mathématique qui se veut être une synthèse des deux approches précédentes. La prise de décision en statistique mathématique s'appuie à la fois sur les observations statistiques que sur les résultats fournis par le modèle probabiliste. En particulier, ce modèle ne postule plus la symétrie des dés, on admet que le dé soit déformé et que la probabilité d'apparition des faces n'est plus la même pour toutes les faces. On utilise ici la définition naïve de la probabilité d'apparition du 6 comme la limite de la fréquence d'apparition du 6. Le problème est qu'il faut que le nombre de lancers soit suffisamment grand. À l'aide de calcul probabilistes, utilisant le théorème «Central Limite», on peut déterminer un intervalle de confiance à 95% pour l'estimation de la probabilité de non apparition du 6 en quatre lancers. De manière approchée on trouve que cette probabilité p est telle que:

$$0,52 - 0,10 < p < 0,52 + 0,10$$

Les observations issues des cent parties nous permettent seulement de conclure que la probabilité de l'événement «non-six» peut être un nombre quelconque de l'intervalle [0,42 ; 0,62], c'est à dire qu'elle peut être aussi bien inférieure à 0,5, auquel cas il vaut mieux miser sur l'apparition du 6, que supérieure à 0,5, et dans ce cas là il vaut mieux miser sur la non apparition du 6.

On ne peut conclure quel est l'événement le plus probable, on ne peut, par exemple, pas détecter si les dés sont pipés. L'attitude raisonnable, dans le cadre d'un raisonnement de statistique mathématique est d'attendre que le nombre d'expériences soit suffisamment grand pour que l'intervalle de confiance pour la probabilité de non apparition du 6 soit situé entièrement à gauche ou entièrement à droite de la valeur 0,5.

Cet exemple illustre le rôle et la vocation des méthodes probabilistes et de la statistique mathématique ainsi que ce qui relie les deux. La théorie des probabilités offre un ensemble de modèles mathématiques permettant de décrire les lois régissant des événements ou le comportement de systèmes fonctionnant sous l'effets de facteurs aléatoires. La statistique mathématique permet de sélectionner dans un ensemble de modèles probabilistes celui qui correspond le mieux, dans un certain sens, aux données dont on dispose.

La situation présentée dans cet exemple est exactement celle que l'on observe dans le cadre des sondages, sondages d'opinion entre autres, les instituts de sondage en France ont une présentation purement descriptive et donnent le pourcentage de manière brute, sans construction d'intervalles de confiance. Dans la terminologie des sondages, l'intervalle de confiance s'appelle une fourchette. Dans le cadre des sondages, le but d'une approche de type statistique mathématique est d'arriver à construire des fourchettes. Grossièrement avec 100 observations ou 100 personnes sondées on a avec une confiance égale à 95%, des intervalles de 10% de part et d'autre de la valeur estimée. En Grande Bretagne les instituts de sondage sont obligés de donner les intervalles d'estimation, il semble que petit à petit cela devienne une norme en France aussi.

19 Un modèle simple la régression linéaire simple.

Le modèle statistique le plus simple à utiliser et à comprendre est le modèle de régression linéaire simple. C'est le modèle qui permet de mesurer l'influence d'une variable quantitative sur une autre variable quantitative. On peut s'intéresser par exemple à l'influence de la température sur la longueur d'une tige en métal, tel qu'un rail de chemin de fer. Tout le monde sait que les rails de chemins de fer ne sont pas placés avec une jointure parfaite en raison de l'élongation due à la chaleur. On peut s'intéresser à l'influence de la taille sur le poids, pour des animaux ou des êtres humains, ou de l'ancienneté sur le salaire. Attention dans les trois exemples que j'ai pris, j'ai pris soin de donner que des variables continues, ou pour lesquelles la continuité est acceptable. J'ai aussi pris soin de prendre trois exemples dans lesquels il n'y a pas vraiment d'ambiguïté sur le sens de la liaison entre les variables, personne ne penserait à dire que c'est la hauteur du mercure dans un tube qui influence la température ou qu'une augmentation de salaire fait augmenter l'ancienneté. La plupart du temps, les utilisateurs pensent que pour mesurer une liaison entre deux variables quantitatives il faut calculer un coefficient de corrélation linéaire. Il est sur que la régression linéaire simple et le calcul du coefficient de corrélation linéaire ont un lien très fort entre eux, mais ont certaines différences qu'il s'agit de comprendre. Tout d'abord il faut noter que pour les variables quantitatives qui ne sont pas continues, on utilise très souvent les coefficients de corrélation et la régression linéaire et qu'il s'agit dans les deux cas d'une approximation. La continuité parfaite des variables n'existe pas très souvent dans la réalité. Une fois cela noté, la différence essentielle, réside dans le fait que dans la corrélation le rôle des deux variables est parfaitement interchangeable, dans la régression linéaire il ne l'est pas. Ce ne sont pas les statistiques qui peuvent dire si une variable est explicative ou à expliquer. Les avantages de la régression linéaire simple par rapport à la corrélation linéaire sont nombreux. Les principes de la régression linéaire simple se généralisent à la régression non linéaire, quadratique ou logarithmique par exemple, mais surtout à la régression multiple, puis à la régression sur variables qualitatives que l'on appelle analyse de variance, et à toutes les variantes de ce que l'on appelle les modèles linéaires, analyses de covariances et autres. Un certain nombre des principes de la régression se retrouvent aussi dans les modèles linéaires généralisés, régression logistique, modèle log linéaire etc. et dans toute la statistique paramétrique.

20 Un jeu de données.

On a mesuré sur 25 sportifs amateurs de Toulouse ou sa banlieue, faisant de la course à pieds ou du cyclisme la fréquence cardiaque dite fréquence seuil qui s'obtient après trois quarts d'heure d'effort soutenu et qui se mesure avec un cardio-fréquence-mètre. C'est une petite machine qui vaut entre 300 et 400 francs et qui est très facile d'utilisation. La question que l'on s'est posée est de savoir si l'âge des sportifs a une influence sur la fréquence seuil.

Voici les données, dans la première colonne l'âge des sportifs dans la deuxième colonne la valeur de la fréquence seuil:

```
30 175
54 165
29 169
51 172
36 170
41 170
40 167
50 165
61 165
23 170
49 170
30 167
32 177
22 169
22 172
32 173
44 168
34 169
32 170
20 172
46 175
45 168
38 172
60 165
58 172
```

21 Phase exploratoire préliminaire

Après avoir insisté sur l'utilisation des probabilités en statistique, il me semble nécessaire d'affirmer que la démarche exploratoire est nécessaire et préalable à toute tentative de modélisation. Si une variable doit être mise dans une analyse de type régression ou analyse de variance il est nécessaire d'étudier sa distribution. Il est donc vivement recommandé de faire un tri simple sur toutes les variables qualitatives, ce que l'on appelle parfois un tri-à-plat, pour les variables quantitatives, on peut calculer les paramètres statistiques usuels, moyenne, écart type, médiane, maximum, minimum, et je recommande chaleureusement l'utilisation des boîtes de dispersion, appelées boxplot en anglais et parfois en français boîtes à moustaches ou boîtes à pattes selon les logiciels. La phase exploratoire est nécessaire pour détecter les erreurs de saisie de données ou les erreurs de codages, mais aussi les valeurs dites aberrantes. Une compréhension ou une interprétation des valeurs aberrantes est un préalable aux traitements plus sophistiqués. Les valeurs aberrantes peuvent avoir une trop grande importance dans un modèle statistique. Mais la démarche exploratoire a ses limites, la vision probabiliste des statistiques n'est pas là pour observer des données mais pour prédire des valeurs à partir de données observées, si on fait de la modélisation c'est pour interpoler, extrapoler; c'est typiquement la situation du sondage. Il faut noter que souvent même si on l'ignore consciemment, on fait de l'extrapolation à partir de méthodes qualifiées d'exploratoires.

Voilà ce que donne une simple commande de SPSS sur la variable fréquence seuil.

Description statistique simple de la variable «fréquence au seuil»:

```
 Nombre d'observations    25


 Moyenne     169.88
 Médiane     170
 Minimum     165
 Maximum     177
 Range        12
 Variance     10.6933
 Écart Type    3.2701
 Std Err      .6540

Boîte de dispersion

           |
       180 +
           |
           |    --+--
           |      |
           |    +-+-+
           |    | * |
           |    |   |
           |    +-+-+
           |    --+--
           |
           |
       160 +
           |
           |
           |
           |
           +-----------------
```

Une version graphique de ce diagramme a été mise à la fin de ce document dans sa version postscript. On remarque qu'il n'y a pas de valeurs aberrantes et que la distribution n'est sans doute pas éloignée de la normalité. Un autre graphique, non présenté ici, l'a confirmé.

22 Mise en liaison des deux variables.

Pour visualiser les liens entre deux variables on peut faire un diagramme de dispersion. Le diagramme de dispersion est une représentation très simple de deux variables. Les axes dans la représentation sont habituellement nommés, axe des x pour l'axe horizontal et axe des y pour l'axe vertical. Implicitement lorsqu'on représente deux variables de cette manière, on a déjà l'idée que la variable que l'on met sur l'axe des y dépend de celle que l'on met sur l'axe des x. On voit par là que la représentation graphique est en quelque sorte une pré modélisation, même si on ignore les terminologies de variables explicatives et de variables à expliquer. Je rappelle que dans la terminologie du collège, la position d'un point sur l'axe des x s'appelle l'abscisse et la position sur l'axe des y l'ordonnée. Ce mode de représentation est très semblable à celui de la géographie qui repère les points par la longitude et la latitude. Pour des raisons de place le diagramme est représenté sur la page suivante.

Diagramme de dispersion de la fréquence au seuil contre l'âge

```
         ++----+----+----+----+----+----+----+----+----+----+----+----+----++
      178+                                                                  +
         |                                                                  |
         |                           *                                      |
         |                                                                  |
         |                                                                  |
      176+                                                                  +
         |                                                                  |
         |                         *               *                        |
         |                                                                  |
         |                                                                  |
      174+                                                                  +
f        |                                                                  |
r        |                           *                                      |
e        |                                                                  |
q        |                                                                  |
u     172+               * *               *            *      *            +
e        |                                                                  |
n        |                                                                  |
c        |                                                                  |
e        |                                                                  |
      170+                  *        *   *    *       *                     +
a        |                                                                  |
u        |                 *      *    *                                    |
         |                                                                  |
s        |                                                                  |
e     168+                                       **                         +
u        |                                                                  |
i        |                         *         *                              |
l        |                                                                  |
         |                                                                  |
      166+                                                                  +
         |                                                                  |
         |                                             *   *     **         |
         |                                                                  |
         |                                                                  |
      164+                                                                  +
         ++----+----+----+----+----+----+----+----+----+----+----+----+----++
          5        15        25        35        45        55        65

                                         âge
```

Chaque étoile sur ce graphique représente un des sportifs de l'échantillon, c'est une autre façon de représenter le tableau de données. Il s'agit de ce que l'on peut appeler une représentation cartésienne, pour chacune des étoiles l'abscisse donne l'âge et l'ordonnée donne la valeur de la fréquence seuil. À partir de ce graphique on peut constater une tendance à la baisse de la fréquence seuil au fur et à mesure que l'âge augmente. Ce diagramme de dispersion est aussi représenté sur la figure 2 en fin de document.

Le but de la régression linéaire est de trouver à partir de ce nuage de points, une droite qui le résumerait au mieux. Si on appelle x, l'âge et y la fréquence au seuil, on cherche une droite d'équation $y = ax + b$ telle que celle ci passe le plus près possible des points du nuage. Le but de la régression est donc de trouver les paramètres a et b qui caractérisent la droite. Contrairement à ce qui est le cas traditionnel en mathématique, on appelle x et y les observations, et ce sont a et b qui désignent les inconnues. J'ai mentionné que Viète qui a introduit les lettres en mathématiques, avait employé les lettres a, b, c comme inconnues et que par la suite ce sont x, y et z qui se sont imposées, en statistique on désigne souvent les quantités inconnues par les premières lettres, mais souvent de l'alphabet grec, $alpha$ ou $beta$.

Sur ces données le calcul de la régression linéaire simple donne $a = 174,22$ et $b = -0,111$, c'est à dire que la droite d'équation:

$$y = 174,22 - 0,111x$$

est la droite la plus proche des points du diagramme de dispersion. Si on accepte cette droite comme représentant les données on peut dire que la fréquence seuil vaut 174 à la naissance et baisse d'environ 1 point tous les 10 ans. Les données sont presque toujours avec des nombres entiers.

23 Régression linéaire, point de vue inférentiel

Il est dangereux avec les données observées de considérer la droite de régression comme un modèle valide, en effet beaucoup de point observés se trouvent assez loin de la droite de régression. L'approche inférentielle doit nous permettre de valider le modèle.

Dans le cadre de la modélisation probabiliste, on considère que les observations des valeurs sont le résultat d'un phénomène aléatoire. On dit, en terme probabiliste que l'observation est la valeur prise par une variable aléatoire, un objet mathématique difficile à définir rigoureusement dans le cadre d'un tel exposé, mais dont le but est de définir l'ensemble des possibilités pour les valeurs qui auraient pu être prises par l'observation. Si on admet, compte tenu de la boîte de dispersion, que c'est la loi normale qui décrit ces possibilités, cela revient à dire que nous avons schématiquement la relation suivante:

$$y = a * x + b + \epsilon$$

Où ϵ désigne, si l'on peut dire le hasard. On dit que ϵ représente une loi normale de moyenne nulle. Ce que cela change, essentiellement par rapport à la vision descriptive c'est que ce que l'on obtiendra ce n'est pas la valeur de a ou b mais une estimation de la valeur de a ou b. Concrètement on obtient une valeur pour a ou b mais aussi une mesure de l'incertitude qui existe dans cette estimation. Cette mesure de l'incertitude dans l'estimation des paramètres, c'est ainsi que l'on appelle a et b, s'appelle «l'erreur standard». Classiquement on obtient aussi dans la régression un test de Fisher pour indiquer si le paramètre a est significativement différent de 0. Dans le calcul du test interviennent un certain nombre de concepts qui reviennent systématiquement dans tous les modèles.

La valeur prédite par la régression est la valeur théorique, ou la valeur attendue, ou la valeur ajustée, si le modèle est bon, dans le cas d'une régression linéaire simple c'est tout simplement la valeur sur la droite. C'est à dire dans l'exemple des données présentées plus haut pour un sportif âgé de 50 ans, c'est la fréquence seuil obtenue par le calcul:

$$\text{fréquence théorique} = 174,22 - 50 * 0,111 = 168,67$$

Le résidu pour une observation c'est l'écart entre la valeur observée, et la valeur théorique.

$$\text{résidu} = \text{valeur observée - valeur théorique}$$

Pour le seul individu de 50 ans de l'échantillon la fréquence vaut 165 le résidu vaut $165 - 168,67 = -3,67$

La somme des carrés des résidus s'appelle la somme des carrés résiduelle. C'est une mesure de l'adéquation entre le nuage et la droite, dans le cas de la régression linéaire simple, plus généralement cela mesure, globalement l'écart entre les valeurs prédites par le modèle et les observations.

Plus on met de paramètres dans un modèle et plus le modèle est complexe. Un modèle complexe a plus de chance de s'approcher des données, si on ne tient compte que de la mesure globale de l'écart entre les observations et les données, mentionnée plus haut on risque d'avantager le modèle le plus complexe. Si on ne dispose que de deux observations, et que l'on fasse une régression linéaire simple on est sur que la droite passera par les deux points est l'ajustement du modèle aux données sera parfait. Il existe un concept important en modélisation qui permet d'éviter ce genre de pratiques, c'est celui de degré de liberté. Le degré de liberté associé à la somme résiduelle des carrés est grosso modo égal aux nombres d'observations moins le nombre de paramètres dans le modèle.

degrés de liberté = nombre d'observations - nombre de paramètres

Dans la régression le nombre de paramètres du modèle est deux, a et b. On divise les sommes de carrés par les degrés de liberté associés, un critère simple pour dire si un modèle est meilleur qu'un autre est d'avoir le plus petit rapport entre la somme de carrés résiduelle et le nombre de degrés de liberté résiduels.

24 Lecture des résultats de SPSS.

Voici ce qu'a donné une commande de régression linéaire sur les données présentées plus haut.

```
-> CURVEFIT /VARIABLES=freq  WITH age
->   /CONSTANT
->   /MODEL=LINEAR
->   /PRINT ANOVA
->   /PLOT FIT.

MODEL:  MOD_1.

Dependent variable.. FREQ              Method.. LINEAR

Listwise Deletion of Missing Data

Multiple R           .41707
R Square             .17395
Adjusted R Square    .13803
Standard Error      3.03600

            Analysis of Variance:

               DF   Sum of Squares      Mean Square

Regression      1         44.64213        44.642127
Residuals      23        211.99787         9.217299

F =       4.84330       Signif F =  .0381

-------------------- Variables in the Equation --------------------

Variable                   B         SE B       Beta          T  Sig T

AGE                -.110907      .050395   -.417071     -2.201  .0381
(Constant)       174.223103     2.064766                 84.379  .0000
```

La syntaxe nous indique que l'on a essayé «d'expliquer» la fréquence cardiaque par l'âge des sportifs. Les coefficients a et b sont dans la colonne B dans le tableau intitulé «Variables in the Equation», il s'agit des estimations des paramètres. La ligne où il y a marqué Constant entre parenthèse représente le coefficient que j'ai appelé b, ce coefficient est parfois appelé intercept, parce que c'est la valeur qui correspond à la valeur 0 de la variable explicative et que cela correspond à l'intersection entre la droite de régression et l'axe des y. La valeur du coefficient que j'ai noté a se trouve sur la ligne qui contient le nom de la variable explicative, AGE en l'occurrence. Il est à noter que quelque soit l'approche, descriptive ou modélisatrice, les valeurs sont exactement les mêmes, je rappelle que ici il s'agit de ce que l'on appelle en statistique des estimations pour les paramètres. La colonne suivant l'estimation des paramètres est celle donnant l'erreur standard liée à l'estimation des paramètres. À partir de la valeur de l'estimation et de l'erreur standard correspondant, on peut construire un intervalle de confiance pour les paramètres. On peut aussi, et c'est ce que fait SPSS, construire un test de Student pour tester si les paramètres sont significativement différents de 0. Dans le cas de la régression linéaire simple le test pour savoir si a est non nul est fait de deux manières, avec le F de Fisher et avec le T de Student, le F est exactement le T élevé au carré, le niveau de significativité est, fort heureusement, rigoureusement le même. Le tableau de lecture d'une régression et d'une analyse de variance sont étrangement semblables.

25 Graphiques.

Voici regroupés en fin de document les graphiques qui permettent d'illustrer la régression linéaire simple. La première figure représente la boîte de dispersion de la variable fréquence au seuil. La deuxième figure représente le diagramme de dispersion de la variable fréquence au seuil sur la variable âge. La troisième figure représente ce même graphique de dispersion avec la droite de régression linéaire simple, obtenue par une approche descriptive. La quatrième figure représente le diagramme de dispersion avec la droite de régression et des courbes correspondant à des intervalles de confiance à 95% de sécurité. La première figure correspond à une démarche statistique exploratoire, la deuxième à une démarche descriptive, qui suppose, toutefois une mise en liaison entre deux variables laisse déjà supposer une idée de modélisation. La troisième correspond à un calcul sophistiqué mais qui ne fait pas intervenir de probabilités. La quatrième figure correspond à une démarche de statistique mathématique ou de modélisation qui inclue un raisonnement probabiliste.

Pour des raisons de commodité de présentation les figures sont chacune sur une page au format paysage.

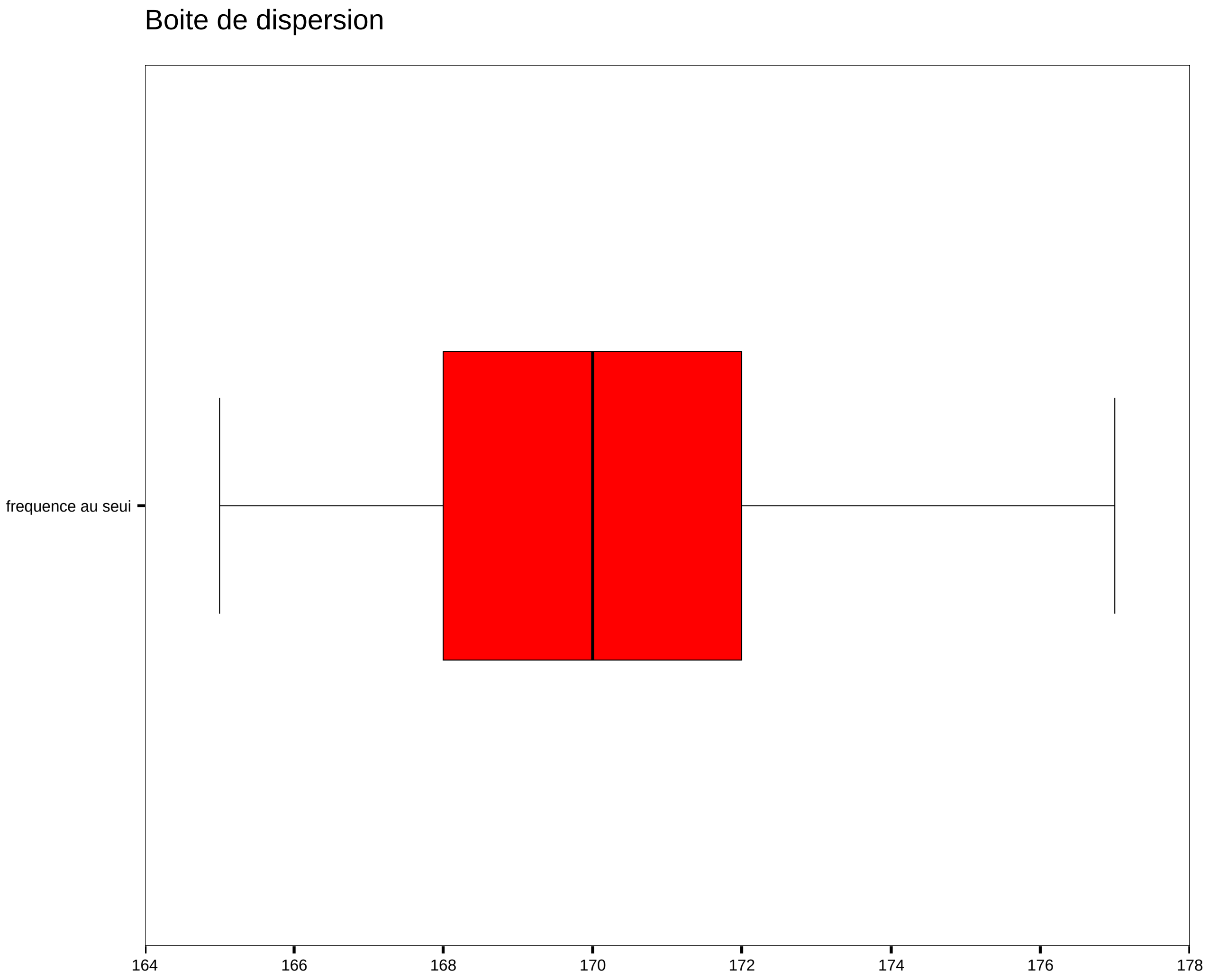
Boite de dispersion
frequence au seui
164
166
168
170
172
174
176
178

FIG. 1 – .

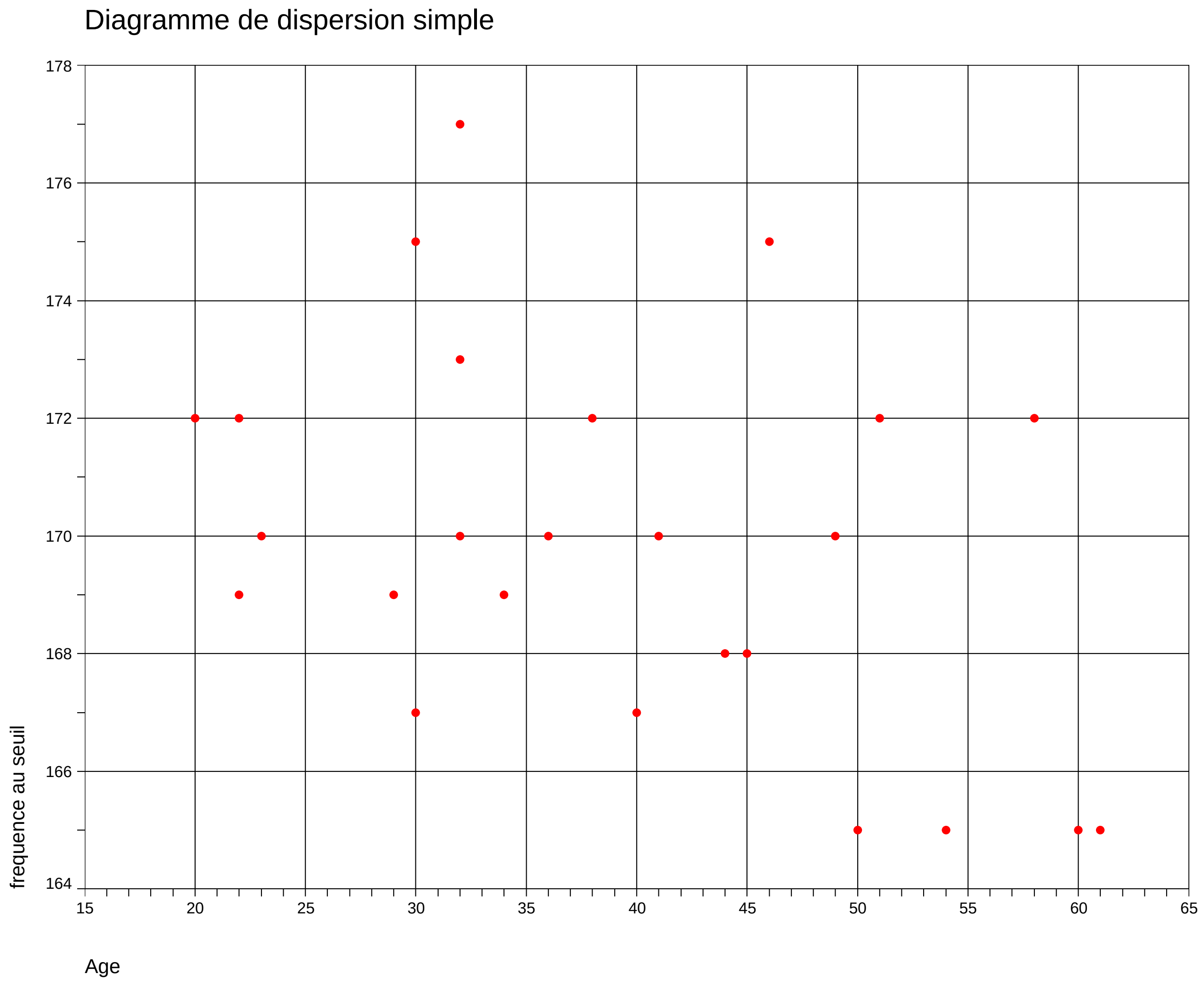
Diagramme de dispersion simple
frequence au seuil
Age
178
176
174
172
170
168
166
164
15
20
25
30
35
40
45
50
55
60
65

FIG. 2 – .

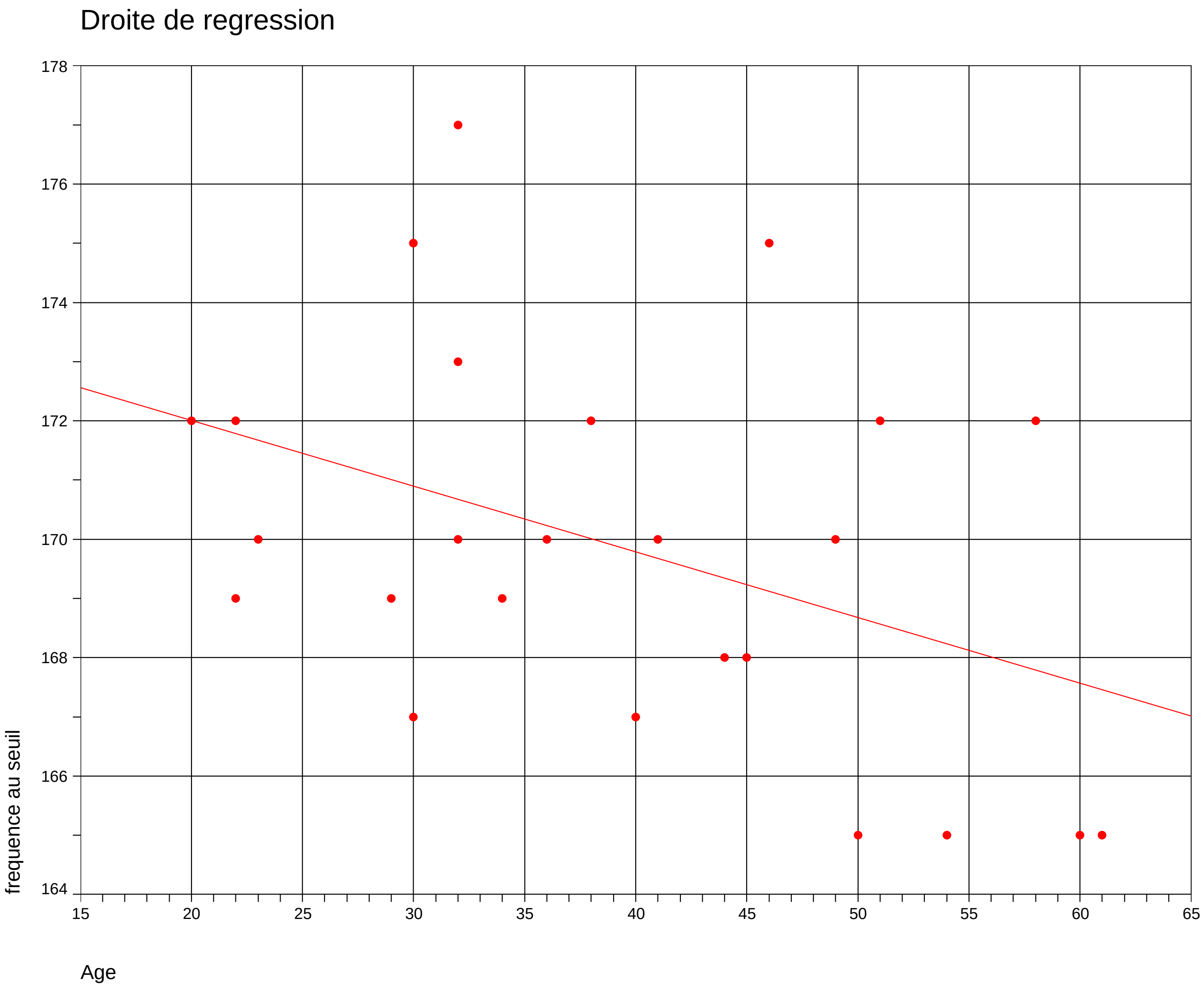
Droite de regression
178
176
174
172
170
168
166
164
15
20
25
30
35
40
45
50
55
60
65
Age
frequence au seuil

FIG. 3 – .

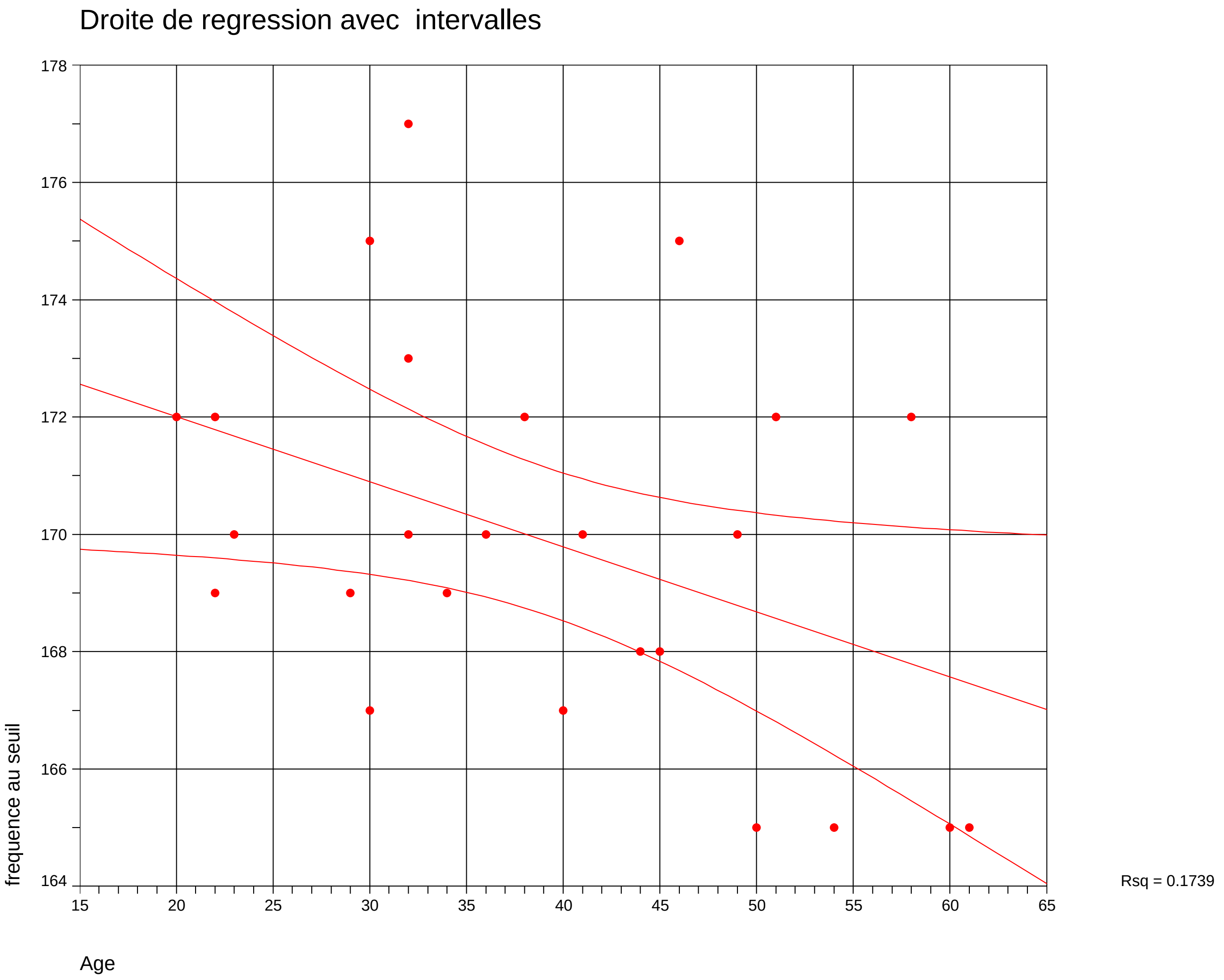
Droite de regression avec intervalles
178
176
174
172
170
168
166
164
frequence au seuil
15
20
25
30
35
40
45
50
55
60
65
Age
Rsq = 0.1739

Fig. 4 – .

26 Commentaires sur ce document

Le document que j'ai écrit pour cet exposé a été rédigé très rapidement. Il a été rédigé en fonction des remarques que j'ai reçues suite à l'exposé du 16 janvier 1998. J'avais envisagé initialement de faire un cours «classique» sur la régression et les modèles statistiques s'adressant à un public relativement averti. Le décalage entre le public attendu et le public observé m'a conduit à une démarche de type modélisation et à essayer autre chose. Durant l'année universitaire 1987-1988 j'ai été assistant en mathématiques à l'Université de Toulouse le Mirail et j'ai assuré des enseignements de statistique à des étudiants d'histoire et géographie, j'avais préparé pour la première séance, une présentation de l'histoire des statistiques et du calcul. En plongeant dans ma mémoire, j'ai repris et actualisé cet exposé pour écrire la partie historique que j'ai mise dans ce document. Je sais très bien que public d'une telle conférence n'est pas nécessairement constitué d'amateurs d'histoire, mais d'après moi c'est un effort fondamental que devrait faire tout scientifique, que de s'intéresser à l'histoire de son sujet de recherche et de sa discipline. Faire l'histoire d'un sujet de recherche ce n'est d'ailleurs ni plus ni moins que la première étape indispensable au démarrage de tout travail de recherche, c'est à dire faire une bibliographie sérieuse. Il va sans dire que cet exposé ne satisfait aucunement à ce critère et ne peut donc être pris comme un document de référence. Un certain nombre de considérations écrites dans ce document sont le produit d'une réflexion superficielle et j'en ai parfaitement conscience. Je reconnais que la rédaction de ce texte est en partie liée à l'actualité immédiate et aux livres que je lis en ce moment ou que je viens de lire.

Ce document ne doit pas être pris pour une introduction rigoureuse aux statistiques et probabilités, mais plutôt comme une étape de sensibilisation à l'intérêt de l'utilisation d'une démarche statistique fondée sur la prise en compte du hasard. Dans le titre de l'exposé, il y a le qualificatif de primaires et en quelque sorte c'est ce qui convient pour décrire ce que j'ai tenté d'écrire, en pensant à un public éloigné des mathématiques et parfois même légèrement allergique à celles ci. Je pense que pour aller plus loin dans la compréhension des statistiques il est indispensable de dépasser le niveau des mathématiques que j'ai présentées ici mais c'est hors de question de faire cela dans le cadre d'un séminaire. Ce que j'ai écrit est plus un essai informel de sensibilisation à la statistique, à ne donc pas prendre trop au sérieux...

Je n'ai pas mentionné, dans le séminaire du 18 Janvier, ni dans celui d'aujourd'hui, les liens existant entre mathématiques et littérature, quoique ces séminaires se tiennent dans une université où la littérature est essen-

tielle. Bien sûr on peut penser à la science fiction, qui est un genre littéraire dans lequel la science joue un rôle important, par définition. Un auteur comme Isaac Asimov, en plus d'être un romancier était un scientifique et un excellent vulgarisateur des sciences. Si la science peut intervenir dans les romans, les romans peuvent aussi inspirer des scientifiques, certaines idées scientifiques sont parfois apparues dans des romans d'anticipation, on pense facilement aux romans de Jules Verne. Certains auteurs modernes ont poussé parfois assez loin l'introduction des mathématiques dans le roman ou dans la construction du roman, je pense bien sûr à l'OULIPO (Ouvroir de Littérature Potentielle) et surtout au génial Georges Perec, trop tôt disparu. J'ai aussi un intérêt pour les romans « scientifiques» d'Umberto Eco, professeur de linguistique à l'université de Bologne, «L'île du jour d'avant» est consacré au problème de la détermination de la longitude, entre autres choses, «Le pendule de Foucault» est comme son titre l'indique son titre consacré à un problème scientifique, mais c'est dans «Le nom de la rose» que j'ai trouvé un chapitre entier qui tourne autour d'une démarche proche de celle que je crois être celle du statisticien, je cite ici un extrait: « Mais alors, osai je commenter, vous êtes encore loin de la solution...
-J'en suis très près, dit Guillaume, mais je ne sais pas de laquelle.
-Donc, vous n'avez pas une seule réponse à vos questions?
-Adso si tel était le cas, j'enseignerais la théologie à Paris.
-À Paris, ils l'ont toujours, la vraie réponse?
-Jamais, dit Guillaume, mais ils sont très sûrs de leurs erreurs.
-Et vous, dis je avec une infantile impertinence, vous n'en commettez jamais d'erreurs?
-Souvent, répondit-il. Mais au lieu d'en concevoir une seule, j'en imagine beaucoup, ainsi je ne suis l'esclave d'aucune. »

Il va s'en dire que j'expose des idées qui sont les miennes sur la façon de faire des mathématiques appliquées et avec lesquelles je comprends très bien qu'on puisse être en désaccord. Au travers des statistiques j'exprime aussi des opinions sur des aspects non mathématiques, il va de soi que si j'ai choisi de parler, dans mon exposé du 16 Janvier 1998, de la peine de mort cela correspond aussi à des positions philosophiques, sociales, politiques etc. avec lesquelles aussi je conçois facilement qu'on puisse être en désaccord. J'accepte les critiques et j'ai cru comprendre que la structure du séminaire ne facilite pas vraiment la discussion. Si vous avez des critiques à faire sur les deux exposés, je vous invite vivement à me les faire connaître. Si la peur d'être enregistré vous empêche de vous exprimer oralement pendant le séminaire, vous pouvez utiliser les adresses courrier et courrier électronique que j'ai mises en-tête du document pour me les faire parvenir. J'en tiendrai compte pour un éventuel séminaire suivant.

Je donne quelques pistes bibliographiques, mais il est évident que j'en oublie beaucoup, il est maintenant très facile, pour qui sait utiliser soigneusement Internet, de trouver des documentations sur les logiciels de statistique et même sur les méthodes, des cours entiers, je conseille la page du Laboratoire de Statistique et Probabilités http://www-sv.cict.fr/lsp/ pour commencer. Si j'ai un peu utilisé Internet pour rédiger ce document et faire quelques copier-coller j'ai aussi utilisé certains documents, non cités dans la bibliographie, mais classiques et indispensables, notamment les dictionnaires, Larousse, Robert 1 et 2, Hachette..., encyclopédie Universalis. Bien évidemment aussi je me suis inspiré des cours que j'ai vu passer depuis fort longtemps.

Références

[1] S. Aïvazian, I. Énukov, L. Méchalkine, *Eléments de modélisation et traitement primaire des données, MIR, Moscou,1986*

[2] V. Pugachev, *Théorie des probabilités et statistiques mathématiques, MIR, Moscou, 1982*

[3] Duncan Cramer, *Introducing statistics for social research, step-by-step calculations and computer techniques using SPSS, Routledge, 1994*

[4] Bernard Bru, *Petite histoire des probabilités, Fragments d'histoire des mathématiques, Brochure A.P.M.E.P. 41, 1981*

[5] SPSS, *SPSS 6.1 Syntax Reference Guide, SPSS Inc., 1994*

[6] Lancelot Hogben, *Les mathématiques pour tous, Payot, 1962*

[7] Philip J. Davis, Reuben Hersh *L'univers mathématique, Gauthiers-Villars, 1985*

[8] Philip J. Davis, Reuben Hersh *L'empire mathématique, Gauthiers-Villars, 1988*

[9] Anders Hald, *A history of probability & statistics and their applications before 1750, John Wiley, 1990*

[10] Laurent Schwartz, *Un mathématicien aux prises avec le siècle, Odile Jacob, 1997*

[11] Stephen Jay Gould, *La mal-mesure de l'homme, Odile Jacob, 1997*

[12] Stephen Jay Gould, *L'éventail du vivant, Le Seuil, 1997*

[13] S.M. Stigler, *The history of statistics: the treatment of uncertainty before 1900. Harvard Univ. Press, 1986*

[14] Colin Ronan, *Histoire mondiale des sciences, Le Seuil, 1988*

[15] Henri Ramet, *Histoire de Toulouse, Le Pérégrinateur, 1994*

[16] Philippe Wolff, *Histoire de Toulouse, Privat, 1994*

[17] John M. Burney, *Toulouse et son université, facultés et étudiants dans la France provinciale du dix neuvième siècle, Presses Universitaires du Mirail, 1989*

[18] Pierre Thuillier, *D'Archimède à Einstein, Les faces cachées de l'invention scientifique, Fayard, 1988*

[19] Pierre Thuillier, *Les passions du savoir essais sur les dimensions culturelles de la science, Fayard, 1988*

[20] Isabelle Stengers, *L'invention des sciences modernes, La découverte, 1993*

[21] Denis Guedj, *La gratuité ne vaut plus rien et autres chroniques mathématiciennes, Le Seuil, 1997*

[22] John J O'Connor and Edmund F Robertson, *the MacTutor History of Mathematics archive, http://www-groups.dcs.st-and.ac.uk/history/, Internet, 1998*

[23] Christian Nitschelm , *Cours d'histoire de l'astronomie, http://www.u-bourgogne.fr/c.nitschelm/histoire.html, Internet, 1998*

[24] Jan Kowalski, *Roman Board Games, http://www.personal.psu.edu/users/w/x/wxk116/roma/rbgames.html, Internet, 1998*

[25] Justin du Coeur, *Medieval and Renaissance Games, http://www.inmet.com/ justin/game-hist.html, Internet, 1998*

Table des matières

1 Préambule, retour sur le séminaire précédent. 2

2 Tentative de réponse à une question. 2

3 Autre retour, la triangularité. 3

4 Droit et mathématiques, équité et égalité. 4

5 Droit et mathématiques, un exemple local. 5

6 Les probabilités, une place à part dans les sciences. 7

7 Fragments d'histoire des sciences, le hasard. 8

8 Préhistoire des probabilités. 9

9 Une énigme de l'histoire des sciences. 10

10 Un autre type d'explication. 12

11 Exemple de raisonnement du type modélisation. 13

12 Mais qu'est ce qu'un «bon» modèle? 15

13 La revanche de Platon? 17

14 Digressions sur la place des mathématiques. 18

15 Les modèles dits déterministes et leurs limites. 20

16 Astronomie, mathématiques, probabilités et statistiques 21

17 La précision, source d'incertitude. 25

18 Raisonnement probabiliste en statistiques. 26

19 Un modèle simple la régression linéaire simple. 28

20 Un jeu de données. 29

21 Phase exploratoire préliminaire 30

22 Mise en liaison des deux variables. 33

23 Régression linéaire, point de vue inférentiel **35**

24 Lecture des résultats de SPSS. **38**

25 Graphiques. **40**

26 Commentaires sur ce document **45**

Bibliographie **48**