COMMISSION DE REFLEXION SUR L'ENSEIGNEMENT DES MATHEMATIQUES

STATISTIQUE

Rapport adopté le 15 mars 2003

Composition de la seconde commission mise en place le 10 mars 2001 à la demande de Jacques Lang, Ministre de l'Education Nationale

Président :

Jean-Pierre KAHANE, professeur émérite à l'Université Paris-Sud, membre de l'Académie des Sciences, président de la CREM

Membres :

- Michèle ARTIGUE, professeur à l'Université Denis Diderot, directrice de l'IREM de Paris 7
- Roger BALIAN, physicien au CEA, membre de l'Académie des Sciences
- Frédéric BONNANS, maître de conférences à l'Ecole Polytechnique, directeur de recherches à l'INRIA
- Rémy COSTE, professeur au lycée Edmond Michelet à Arpajon, membre du CNP
- Claude DESCHAMPS, professeur de mathématiques spéciales au lycée Louis le Grand, membre du CNP
- Catherine DUFOSSE, professeur au lycée de Marseilleveyre à Marseille
- Jean-Claude DUPERRET, professeur à l'IUFM de Reims, responsable du centre IUFM de Troyes
- Yves ESCOUFIER, professeur à l'Université de Montpellier II
- Catherine HOUDEMENT, maître de conférence à l'IUFM de Rouen
- Francis LABROUE, IA-IPR de mathématiques, Académie de Créteil
- Rémi LANGEVIN, professeur à l'Université de Bourgogne
- Michel MERLE, professeur à l'Université de Nice, membre du CNP
- Daniel PERRIN, professeur à l'Université Paris-Sud et à l'IUFM de Versailles
- Antoine PETIT, professeur à l'ENS de Cachan
- Claudine ROBERT, professeur à l'Université Joseph Fourier et à l'IUFM de Grenoble, présidente du GEPS de mathématiques
- Marc ROSSO, directeur du département de mathématiques de l'ENS, rue d'Ulm
- Claudine RUGET, doyenne de l'Inspection Générale de mathématiques.

SOMMAIRE

Introduction

Yves Escoufier
Laboratoire de Probabilités et Statistiques
Université Montpellier 2
yves.escoufier@univ-montp2.fr

Invité à coordonner un rapport sur la Statistique, j'ai proposé aux membres de la commission de réflexion sur l'enseignement des mathématiques de le construire par le rassemblement d'articles courts écrits par différents auteurs. Ce choix voudrait apporter plusieurs éclairages complémentaires sur cette discipline.

D'abord il met en évidence qu'il n'y a pas de pratiques statistiques sans données et qu'il n'y a pas de données sans des questions issues des domaines scientifiques, économiques, sociaux ou industriels qui justifient l' intérêt qu'on leur porte et le plus souvent provoque et organise leur collecte. Les textes qui suivent abordent des problèmes issus de la banque, de l'industrie des cosmétiques, de l'industrie agro – alimentaire, de l'étude des questionnaires ou de la pratique quotidienne de la mesure de différentes grandeurs physiques. On retiendra que dans les démarches qu'ils décrivent, la Statistique n'intervient pas après un raisonnement interne au domaine concerné ; elle ne s'applique pas après comme une fioriture finale ; elle est inhérente à la construction méthodologique ; elle fournit les outils de la structuration des étapes successives et ceux de l'analyse des résultats intermédiaires ou terminaux. Cette interpénétration de la Statistique et des domaines scientifiques, économiques, sociaux et industriels est une spécificité de cette discipline. Elle alimente l'intérêt qu'on peut lui porter puisqu'elle ouvre à ses spécialistes des champs multiples de l'activité humaine. Elle nourrit aussi les critiques que l'on entend à son sujet lorsque ses résultats deviennent des éléments du débat citoyen. Le statisticien a-t–il pris en compte toute la complexité du problème ? Le spécialiste du domaine concerné a–t–il su mettre en œuvre les méthodes statistiques les mieux adaptées ? Le journaliste ou l'homme politique ont–ils une compréhension suffisante des résultats qu'ils commentent ?

Une seconde information apportée par ces articles réside dans la diversité des méthodes mises en œuvre. Des descripteurs simples comme la moyenne et la variance sont utilisés mais les auteurs font aussi appel à des outils plus complexes issus de la Statistique non paramétrique , des méthodes d'analyses multidimensionnelles ou des tests. La panoplie des méthodes statistiques est très riche et elle s'enrichit constamment. Comme toute discipline, elle le fait de manière endogène en s'interrogeant sur les résultats acquis et en essayant de forger des outils qui ne souffrent pas des limitations reconnues aux outils existants. Elle le fait aussi en se confrontant à des problèmes nouveaux donnant naissance à des problématiques et des données d'un type original. Cet affrontement permanent au réel est une motivation incessante pour le statisticien : la recherche n'est jamais finie ; la pratique est toujours renouvelée. Elle est aussi une exigence car elle interdit de considérer qu'on n'a plus à apprendre. On notera aussi dans tous ces articles la place prise par l'informatique. Elle apparaît de façons variées : pour faire des calculs, pour permettre des simulations, pour gérer de grands ensembles de données ou pour la mise en œuvre de certaines méthodes . C'est là un élément incontournable de la Statistique d'aujourd'hui, évident pour les pratiques statistiques, tout aussi nécessaire dans la recherche elle même. Et ceci a bien sûr des conséquences dans la formation des statisticiens.

Un troisième constat peut être fait : même si la plus grande réserve dans l'emploi des mathématiques leur a été recommandée, les différents auteurs font tous appel à des formules et des expressions mathématiques pour décrire ce qu'ils font. Comme j'essaie de le montrer dans le texte que j'ai écrit pour cette annexe, les concepts et le langage des probabilités et donc des mathématiques sont nécessaires à la description des objets de la Statistique et des démarches du statisticien. Certes, dans des cours de découverte de la Statistique, on peut, en particulier en faisant appel à la simulation, faire appréhender certaines notions fondamentales par l'expérimentation. Ceci est important dès lors que tout citoyen doit être préparé à lire un journal ou écouter une radio. Il rencontrera inévitablement des pourcentages, des graphiques, des prévisions. Rechercher par des expériences simples à développer son esprit critique à leur sujet est une nécessité. Mais dès que l'usage des statistiques devient plus intense et plus professionnel, l'écriture mathématique est inévitable. Comme d'autres, je crois que cette constatation justifie que l'enseignement de la Statistique soit confiée dans le secondaire aux enseignants de mathématiques. Bien sûr, il faut les aider à découvrir cette discipline et à savoir y trouver des applications particulières d'objets et de résultats mathématiques plus généraux.

Ces professeurs de mathématiques des lycées ont été désignés comme les lecteurs préférentiels des articles qui suivent. En conséquence, dans chacun des textes, une partie au moins est assez détaillée pour pouvoir être reproduite dans une situation analogue par un lecteur possédant une formation mathématique raisonnable et prêt à manipuler des données. Ce serait une grande réussite pour ce rapport si sa lecture contribuait au développement de travaux de groupes en Statistique ou alimentait la réflexion sur les laboratoires de mathématiques que la commission appelle de ses vœux dans les lycées.

INCERTITUDES DES MESURES DE GRANDEUR

Claudine Robert, Jacques Treiner
claudine.robert@imag.fr, jacques treiner@noos.fr

Vers les années 1960, le livre de physique le plus utilisé en classe de terminale scientifique était le *Cessac et Treherne* à couverture verte et bleue ; un chapitre de cet ouvrage est " Incertitude des mesures et calculs approchés ". Après avoir annoncé que *" mesurer une grandeur, c'est chercher combien de fois elle contient une grandeur de la même espèce choisie comme unité* ", on trouve les paragraphes suivants, dont nous citons les éléments les plus importants :

-Valeur exacte et valeur approchée :

" Le nombre a, résultant de la mesure d'une grandeur A, n'est qu'une valeur approchée de A. Si x est la valeur exacte de A, δa=a-x est appelée erreur absolue de la mesure. "
...
" Les erreurs systématiques sont celles qu'entraîne l'emploi de méthodes ou d'instruments imparfaits. "
" Les erreurs accidentelles sont surtout imputables à l'imperfection de l'opérateur ; contrairement aux précédentes, elles sont commises tantôt en plus, tantôt en moins….Jamais l'expérimentateur le mieux outillé et le plus habile ne peut être sûr d'atteindre la valeur exacte de la grandeur qu'il mesure. "

-Incertitude absolue, présentation du résultat d'une mesure :
" L'erreur absolue n'étant pas connue, on en cherche un majorant Δa, que l'erreur absolue n'atteint probablement pas, mais qu'elle pourrait atteindre dans le cas le plus défavorable, sans toutefois la dépasser. Le résultat de la mesure est alors présenté sous la forme : a ±Δ a. "

-Calculs d'incertitudes :
L'ouvrage cité énonce les théorèmes des incertitudes absolues (l'incertitude absolue d'une somme ou d'une différence est la somme des incertitudes absolues) et relatives (l'incertitude relative sur un produit ou un quotient est la somme des incertitudes relatives*).*

Le texte ci-dessous reprend ces trois points, avec des outils de nature probabiliste, totalement écartés dans la présentation faite ci-dessus. Il est destiné d'une part à donner un aperçu du calcul d'incertitude de mesures tel qu'il se pratique en laboratoire ou en milieu industriel, d'autre part à montrer qu'il convient au minimum d'inclure, au niveau du lycée, la *mesure* dans la panoplie des expériences aléatoires que l'élève rencontre.

Le but est de dégager des méthodes, et on ne détaillera ni les outils théoriques, ni la complexité de nombreuses situations pratiques.

1- Valeur exacte ou valeur approchée

● Une première découverte (désillusion pour certains, progrès de la pensée pour d'autres) est qu'on ne peut en général pas parler de valeur *exacte* de la grandeur à mesurer, sauf dans

certains cas, par exemple si cette grandeur est une constante mathématique. Ainsi, si on veut mesurer π en choisissant n points au hasard dans un carré et en déterminant la proportion de ceux qui sont dans le cercle unité, on peut parler de la valeur exacte de π (définie à partir la limite d'une série par exemple) et de la valeur approchée obtenue par le procédé de mesure choisi.

Prenons quelques exemples d'autres situations où le terme de valeur exacte n'est pas approprié.

La taille d'un individu.
Chez un adulte, cette taille varie d'environ un centimètre entre le lever et le coucher (effet de tassement diurne). La taille dépend donc de la précision demandée pour son évaluation. À un mètre près, la majorité des adultes mesure deux mètres. Au millimètre près, il faut préciser le moment du jour où la mesure est faite.

La largeur d'une table. Une table n'est pas un objet mathématique, c'est une table réelle, dont la " largeur " varie selon l'endroit où on la mesure. Cette variation résulte du processus de fabrication lui-même, mais aussi du vieillissement du bois, qui se contracte ici, se dilate là, et se gauchit. On pourra noter les différentes valeurs mesurées, effectuer leur moyenne, et observer la distribution des valeurs mesurées autour de cette valeur moyenne.
N'oublions pas non plus ici le " paradigme de la longueur des côtes bretonnes " : parler de " largeur de la table " n'a de sens que si l'on est capable d'isoler cet objet de son environnement. Or, si l'on se place à l'échelle moléculaire, c'est la notion-même de *frontière* entre la table et le reste du monde qui disparaît. On passe de façon continue de l'intérieur de la table à l'extérieur (sur une échelle de quelques distances moléculaires), et d'ailleurs si un bois possède une odeur, c'est bien parce que des molécules le quittent sans arrêt. La notion usuelle de " largeur " perd donc son sens en deçà de l'échelle de quelques molécules, ce qui, reconnaissons-le, ne pose pas grande difficulté pour la vie quotidienne. On met là le doigt sur le fait qu'un concept n'est pertinent qu'à une certaine échelle d'appréhension du monde.

La température et la pression. Ce sont, par construction, des grandeurs qui ont une dispersion. La température, par exemple, est proportionnelle à l'énergie cinétique *moyenne* des particules du milieu. Or l'énergie cinétique totale, proportionnelle à une somme de variables aléatoires (le carré des vitesses des particules), est une variable aléatoire, et sa moyenne également. Pour un système macroscopique, la variabilité est inobservable (l'écart-type est en $1/\sqrt{N}$, où N est le nombre de molécules). Elle devient cependant perceptible si l'on diminue le nombre de constituants, comme dans les noyaux atomiques et les petits agrégats moléculaires. A l'échelle d'une particule, le concept de température n'a plus de sens. Où se situe la transition ? Des chercheurs travaillent en ce moment-même sur cette question, en étudiant notamment la signature des transitions de phase connues dans des systèmes de petite taille.

Le nombre d'habitants d'un pays. On peut avoir l'impression qu'il s'agit d'un nombre entier bien défini. Il l'est effectivement, à chaque instant, mais quelle est l'échelle de temps de sa variation ? Il y a sans arrêt des gens qui meurent, disons 600 000 par an en France, à peu près autant qui naissent (un peu plus). Ca fait de l'ordre de 1,2 à 1,3 millions de signaux +1, -1 à distribuer dans l'année. Pour obtenir un ordre de grandeur, supposons que cela se fasse de façon uniforme (il y a des gens qui prétendent que ce n'est pas le cas, et qu'il y a plus de naissances les soirs de pleine Lune, mais ce n'est pas confirmé par l'examen des chiffres dans

les maternités !). Comme il y a environ 30 millions de seconde dans une année, le nombre d'habitants fluctue sur une échelle de 25 secondes. Si l'on trace le nombre d'habitants en fonction du temps, on obtient donc une courbe en dents de scie (avec diverses variations saisonnières, car les naissances et les décès ne se répartissent en réalité pas de façon uniforme !).
Remarquons que dans cette discussion, la question de la *détermination expérimentale* du nombre d'habitants a été laissée de côté. Il est intéressant d'y venir. Le nombre d'habitants à un instant donné existe bien, mais il est cependant impossible à déterminer pratiquement, car le processus de mesure (le recensement) s'effectue sur une échelle de temps bien supérieure à celle de ses fluctuations qui, comme on l'a vu, est de l'ordre de 25 secondes. On est dans un cas où le *temps de réponse* de la mesure est plus lent que le *temps caractéristique des variations* de la grandeur mesurée. Et ce n'est pas tout. Il reste la question du *comptage,* nécessairement entaché d'erreurs, des vraies erreurs cette fois (là, c'est de l'expérimentateur qu'il s'agit). Comme on l'a vu dans les élections américaines de 2001, cela peut conduire à des effets rocambolesques si la décision à prendre requiert une précision plus grande que l'erreur.

Les raies spectrales. Elles ont toujours une " largeur " qui est reliée à la durée de vie d'états excités. On attribue du reste une largeur en énergie aux états eux-mêmes (qu'il s'agisse de l'échelle atomique, nucléaire, ou de l'échelle des particules dites élémentaires).

Notons enfin qu'une dispersion de la grandeur à mesurer peut également résulter de l'influence de paramètres dont on ne contrôle pas la variation : pression ou température lors de la mesure d'un volume, température lors de la mesure d'une résistance, variation temporelle etc.

Insistons donc sur le point illustré ci-dessus : la notion de " vraie valeur " d'une grandeur n'a en général pas de sens et il est préférable de parler de valeur théorique, de référence ou admise : certaines mesures visent ainsi à définir une mesure de référence, d'autres à retrouver une valeur admise (étalonnage d'appareils), d'autres (TP ou expériences de physique) à vérifier certaines prédictions pour des grandeurs calculées à partir de valeurs de références connues (telle la constante de gravitation).
On admettra dans la suite de ce document que la variabilité des résultats de mesure est à rechercher dans les appareils de mesure et/ou chez l'expérimentateur, et que les variations propres de la grandeur à mesurer sont négligeables par rapport aux autres variations mentionnées ci-dessous, ce qui rend pertinent le concept de valeur de référence ou de valeur théorique.

Au passage, relevons enfin l'ambiguïté de la locution *combien de fois* dans la phrase " *mesurer une grandeur, c'est chercher combien de fois elle contient une grandeur de la même espèce choisie comme unité* ",. Dans le contexte du livre cité, un nombre décimal se cache derrière cette locution. Mais la notion mathématique d'incommensurabilité de deux segments, origine de drames pour l'école pythagoricienne, et l'existence de segments incommensurables, tels le coté et la diagonale d'un carré permettent d'assurer que certaines grandeurs ne pourront jamais être mesurées avec exactitude (si on sait mesurer avec exactitude le coté d'un carré, on ne peut pas en faire autant pour la diagonale) : la mesure

donne une valeur approchée et il ne s'agit pas là d'un défaut du processus de mesure qu'il convient de corriger.

● Revenons maintenant à l'instrument de mesure.

Il est caractérisé par
- son *temps de réponse,*
- son *exactitude*, qui se décline en *justesse* (pas d'erreurs systématiques) et *fidélité* (reproductibilité des indications de l'appareil),
- sa *sensibilité.*

Faire une mesure, c'est toujours mettre en interaction un appareil avec le système à étudier, c'est donc enregistrer la *réponse* de l'appareil à une *excitation* produite par le système.
La réponse de l'instrument de mesure met un certain temps à s'établir, c'est le *temps de réponse*. Pour un phénomène indépendant du temps, ce temps de réponse n'est pas important. Pour un phénomène qui varie dans le temps, il faut s'assurer que le temps de réponse de l'appareil est nettement plus petit que l'échelle de variation temporelle de la grandeur à mesurer.

Quelques exemples.

- Une chauve-souris évalue les distances d'obstacles ou de proies par émission-réception d'ultra-sons. Le système n'est efficace que parce que l'intervalle de temps au cours duquel un train d'onde est émis, renvoyé par l'obstacle, reçu par l'animal et décodé par son cerveau est suffisamment bref pour que la position de l'animal pendant ce temps ait peu varié. Sinon, c'est la collision assurée ou l'impossibilité de se nourrir : exit la chauve-souris de la diversité des espèces !
- Certaines jauges de pression fonctionnent par déformation d'une membrane qui constitue l'une des armatures d'un condensateur. La mesure de la capacité de ce condensateur est reliée à la pression exercée sur la membrane. Pour pouvoir suivre des variations temporelles de la pression, le temps de réponse de la membrane (réponse mécanique), doit être petite devant l'échelle de temps de variation de cette pression.
- Lors d'un titrage acide-base, après chaque ajout de réactif titrant, le temps mis pour atteindre le régime permanent d'échange ionique au niveau de l'électrode de verre est bien supérieur à celui de la transformation chimique.

Un appareil de mesure fonctionne bien dans une certaine plage de valeurs de la grandeur à mesurer. Dans la mesure du possible, il faut faire fonctionner un appareil là où sa *sensibilité* est maximale, c'est-à-dire dans un domaine où une faible variation de la grandeur à mesurer produit une variation observable de l'indication de l'appareil.
Dans le cas de la jauge de pression cité plus haut, les limites extrêmes du domaine sont, vers les basses pressions, une déformation de la membrane trop petite pour être mesurée, vers les hautes pression, la limite d'élasticité de la membrane.
Il faut distinguer sensibilité et justesse. Un appareil peut être sensible sans être juste (par exemple s'il est mal calibré). Dans le cas où la grandeur à mesurer a une dispersion intrinsèque négligeable, on dira qu'une mesure est d'autant plus exacte que l'appareil est juste et sa dispersion, faible.

Les constructeurs fournissent des indications concernant la précision de leurs appareils sous forme d'incertitudes à attribuer aux mesures effectuées dans des conditions bien précises. Il faut *se reporter aux notices* de fabrication pour connaître le sens précis… de la " précision " indiquée. Les incertitudes sont de nature très variée. Prenons l'exemple d'une boîte de résistances fournie avec une " précision " affichée de 0,5 %. Cette précision recouvre un aspect d'échantillonnage (le fabricant fabrique des milliers de boîtes dont les résistances varient d'un exemplaire à l'autre), et un aspect de fonctionnement (la résistance change avec la température du fil, qui dépend elle-même de l'intensité du courant qui le parcourt). Le fabricant donne une limite à l'effet de ces différents facteurs sur la valeur des résistances de la boîte, mais il s'agit d'une moyenne et certains appareils font mieux, d'autres moins bien.

● L'opérateur.

La dernière cause de variation des résultats de la mesure d'une grandeur physique réside dans les appréciations de l'opérateur lui-même. On ne refait jamais la mesure exactement dans les mêmes conditions, parce que l'appréciation de l'opérateur change d'une mesure à la suivante : erreur de parallaxe dans le repérage d'un trait de jauge, effets de ménisque dans une pipette, fatigue etc. D'une mesure à l'autre, pour un appareil de précision donnée, le résultat varie. On pourra parler de mesure *juste* si l'opérateur a évité toute *erreur systématique*.

Une mesure comporte en général plusieurs opérations dont chacune peut être source de variabilité. Il est important de savoir distinguer les sources de variabilité importante de celles qui sont négligeables : dans le premier cas, il faudra répéter plusieurs fois l'opération, dans le second cas ce ne sera pas nécessaire. S'il faut, par exemple, prélever un liquide avec une pipette et en effectuer la pesée, la source principale de variabilité sera souvent dans l'utilisation de la pipette : on prélèvera *plusieurs fois* du liquide dont on n'effectuera *qu'une seule* pesée.

2- Présentation du cadre probabiliste du calcul d'erreur.

2-1 : Modèle

Une mesure d'une grandeur est la valeur d'une variable aléatoire X qu'on décompose en somme d'une constante μ (mesure de référence ou mesure théorique de la grandeur, à déterminer ou à estimer) et d'une variable aléatoire ε, d'espérance nulle si la mesure est juste (on se placera ici toujours dans ce cas).

L'écart-type (théorique) σ de X quantifie la dispersion (on ne distingue pas dans ce modèle celle qui est liée à l'instrument et celle qui est liée à l'opérateur).

Autrement dit, on remplace la notion d'erreur accidentelle par celle d'incertitude aléatoire : la variabilité de la mesure n'est pas un " accident " évitable, mais est inhérente au processus de mesure si celui-ci est suffisamment sensible.

La notion de reproductibilité de la mesure signifie qu'on peut associer à la répétition de n mesures un modèle où le résultat $\boldsymbol{x}= (x_1,\ldots,x_n)$ de ces mesures est une réalisation d'un échantillon d'une loi de probabilité. Soit encore :

$$\boldsymbol{x}=(x_1,\ldots,x_n) \text{ est une valeur de } \boldsymbol{X}=(X_1,\ldots,X_n),$$

où les variables X_i sont indépendantes et de même loi . On écrira $X_i=\mu+\varepsilon_i$, où μ est la valeur théorique de la grandeur mesurée.

2-2 Présentation des résultats de *n* mesures.

On peut présenter le résultat des mesures par le couple moyenne, écart-type ($\bar{x}$,s) (sans oublier de spécifier le nombre de mesures faites), avec, pour une série $\boldsymbol{x}=(x_1,\ldots,x_n)$ de mesures :

$$\bar{x}=\frac{x_1+\ldots x_n}{n} \quad \text{et} \quad s=\sqrt{\frac{1}{n}\sum(x_i-\bar{x})^2}$$

Au plan de la modélisation, les n mesures sont considérées comme une réalisation d'un échantillon $X=(X_1,\ldots,X_n)$, d'une loi de probabilité P. Soit μ la moyenne théorique (ou espérance) et σ l'écart-type des variables aléatoires X_i, c'est à dire de la loi P. La loi de probabilité de la variable aléatoire $\bar{X}=\frac{X_1+\ldots X_n}{n}$ est entièrement déterminée par P. Par linéarité de la moyenne, l'espérance de $\bar{X}$ est μ ; comme les variables X_i sont indépendantes, la variance de leur somme est la somme des variances, soit $n\sigma^2$; en divisant la somme des mesures par n, on divise la variance par n^2 ; la variance de $\bar{X}$ est donc σ^2/n, et l'écart-type, encore appelé erreur standard, vaut $\sigma/\sqrt{n}$.

On notera aussi que s est une réalisation de la variable aléatoire $S=\sqrt{\frac{1}{n}\sum(X_i-\bar{X})^2}$.

Exemple

On a 100 mesures du nombre π par la méthode décrite au début du paragraphe 1 ; la moyenne est $\bar{x}$ =3,14, et l'écart-type est s=0,05.

En dehors des calculs théoriques, on peut illustrer que les moyennes d'une série de mesures à une autre fluctuent d'autant moins que la taille de la série est grande et par conséquent la moyenne de plusieurs valeurs est " meilleure " (au sens de sa reproductibilité) que le résultat d'une mesure unique. Le tableau ci-dessous permettent de visualiser les dispersions des moyennes pour de 4,10 ou 50 mesures.

	moyenne	Ecart-type	nombre	min	max
m(1)	3,138	0,045	100	3,033	3,262
m(4)	3,140	0,023	20	3,107	3,197
m(10)	3,140	0,020	20	3,109	3,185
m(50)	3,144	0,006	20	3,133	3,153

La première ligne fournit les résumés pour 100 mesures. La ligne m(i), i=4,10,50 concerne des séries de taille 20, un terme de la série étant la moyenne de i mesures.

Le livre de terminale de Cessac et Treherne, témoigne d'une habitude ancienne et bien ancrée, qui consiste à présenter le résultat d'une série de mesures sous la forme $a \pm \Delta a$. Dans le cas de n mesures, a= $\bar{x}$.

On notera que même si on dispose d'une valeur de référence admise de la grandeur étudiée, c'est à dire si μ est connue, compte-tenu du caractère probabiliste de la mesure, on ne peut plus en général donner un intervalle borné, centré en $\bar{x}$, dont on soit sûr qu'il contienne μ (i.e qui ait une probabilité égale à 1 de contenir μ).

La loi commune des variables X_i est, dans le cas d'incertitudes de mesures, le plus souvent une loi de Gauss. Dans ce cas, pour toute valeur de n, la loi de $\bar{X}$ est une loi de Gauss de même espérance μ et d'écart type $\sigma/\sqrt{n}$. On sait alors calculer simplement, par référence à une loi de probabilité connue (loi de Student), la probabilité $p_n(k)$ que μ soit dans l'intervalle aléatoire :

$$[\, \bar{X} - kS/\sqrt{n-1} \;\; ; \;\; \bar{X} + kS/\sqrt{n-1} \,]^1$$

Les valeurs de $p_n(k)$ sont tabulées. Pour $n>30$, ces valeurs ne varient presque plus en fonction de n et :

$$p_n(1) \,\%\, 0{,}66 \qquad p_n(2) \,\%\, 0{,}95 \text{ et } p_n(3) \,\%\, 0{,}99 \text{ .}$$

Le résultat de n mesures sera en général présenté sous l'une des formes suivantes :

- $\bar{x} \pm s/\sqrt{n-1}$: cette écriture signifie qu'en estimant μ par $\bar{x}$, la précision du résultat est $2\, s/\sqrt{n-1}$ au niveau de confiance 0,66.

- $\bar{x} \pm 2\, s/\sqrt{n-1}$: cette écriture signifie qu'en estimant μ par $\bar{x}$, la précision du résultat est $2\, s/\sqrt{n-1}$ au niveau de confiance 0,95.

- $\bar{x} \pm 3\, s/\sqrt{n-1}$: cette écriture signifie qu'en estimant μ par $\bar{x}$, la précision du résultat est $3\, s/\sqrt{n-1}$ au niveau de confiance 0,99.

Un niveau de confiance $1-\alpha$ signifie qu'on considère une réalisation d'un intervalle aléatoire qui a une probabilité $1-\alpha$ de contenir μ. .

Remarque :

---

[1] Si la loi commune des variables X_i est une loi de Gauss de moyenne μ et d'écart-type σ, alors la loi de $Z = (X-\mu)/(\sigma/\sqrt{n})$ est une loi de Gauss de moyenne 0 et d'écart-type 1 ; on peut montrer que les variables aléatoires Z et S^2 sont indépendantes, que $T = n\,S^2/\sigma^2$ suit une loi du khi-deux à $\nu = n-1$ degrés de liberté et donc que $R_\nu = \dfrac{Z}{\sqrt{T/\nu}}$ suit une loi de Student à ν degrés de liberté. Des tables numériques donnant les probabilités que R_ν soit dans certains intervalles sont dans tous les livres de statistiques. Il s'avère qu'on vient de jongler entre n et $n-1$, et pour une première approche, on pourrait aussi supposer " n grand " et ne pas distinguer $\sqrt{n}$ et $\sqrt{n-1}$.

Les livres calculent souvent, au lieu de s, la quantité $s'=\sqrt{\frac{1}{n-1}\sum(x_i-\bar{x})^2}$ et les intervalles ci-dessus s'écrivent alors :

$$\left[\bar{x}-k\frac{s'}{\sqrt{n}},\bar{x}+k\frac{s'}{\sqrt{n}}\right].$$

Dès que n est grand, la division par n ou n-1 ne change de toute façon pas beaucoup les résultats ! Il convient surtout de se souvenir de l'ordre de grandeur de la précision, à savoir $2s/\sqrt{n}$ pour un niveau de confiance 0,95.

Dans le cas où la loi commune des variables X_i n'est pas une loi de Gauss mais admet cependant une espérance et une variance, on peut utiliser le théorème central limite : pour n grand, la loi de $\overline{X}$ est approximativement une loi de Gauss d'espérance μ et d'écart-type $\sigma/\sqrt{n}$, où σ est l'écart-type de la loi des variables X_i.

Ainsi, pour n grand, on peut à partir de la loi de Gauss, avoir une bonne estimation de la probabilité que μ soit dans un intervalle du type $[\overline{X}-k\sigma/\sqrt{n}\ ;\ \overline{X}+k\sigma/\sqrt{n}\ ]$. On peut estimer σ par s ; le résultat de n mesures est ainsi le plus souvent présenté suivant sous l'une des formes ci-dessus dès que n>30.

Ces considérations ne peuvent pas aujourd'hui être exposées à des élèves de lycée, où on se contente de résumer les mesures par le triplet (moyenne, écart-type, nombre de mesures) ou par $\bar{x}\pm s/\sqrt{n}$.

3- Calculs d'incertitude

En pratique, on ne mesure pas toujours directement la grandeur d'intérêt : celle-ci peut être fonction de grandeurs qui elles sont aisément mesurables. On étudie ici une méthode de calcul d'incertitude pour une situation de ce type. Le principe est de " rester dans le monde Gaussien ", c'est à dire dans un modèle où on manipule des lois de Gauss ; pour cela, on passe par des développements limités autour de l'espérance de la loi de la mesure de la grandeur d'intérêt. Cette méthode est justifiée par le fait que, dans le cadre des erreurs de mesure, l'écart-type de la loi qui modélise les mesures est petit devant son espérance.
Plutôt que de développer le cas général, nous avons choisi ici d'étudier un exemple simple.

● On veut caractériser la forme des feuilles de papier A4 (on suppose qu'on ne connaît pas le procédé qui définit les longueurs et largeurs *théoriques* des feuilles de format A1,A2,A3,A4). Pour cela, on dispose (cf. le tableau ci-dessous) des moyennes et écarts-type de mesures (faites au double décimètre) de longueur et la largeur de 100 feuilles A4, supposées *identiques*.

	Moyenne	Ecart-type	minimum	maximum
largeur	20,96	0,09	20,7	21,2
longueur	29,71	0,10	29,5	30,0

Dire que les 100 feuilles sont identiques, c'est ici faire l'hypothèse que les 100 couples (x_i,y_i) de mesures obtenues sont les valeurs de variables aléatoires indépendantes (X_i,Y_i) et de même loi. On admettra de plus que les mesures de largeur X_i et de longueur Y_i sont indépendantes et suivent des lois de Gauss. On pourra ainsi écrire que x_i et y_i sont une réalisation (i.e ; une valeur) des variables suivantes :

$$X_i = \lambda + E_i \qquad Y_i = L + E'_i$$

où E_i et E'_i sont des variables aléatoires indépendantes suivant des lois de Gauss d'espérance nulle et d'écarts-type respectifs σ et σ' , petits devant λ et L.

En pratique, pour le type d'approximation faits ici, on supposera que :

$$\tau = \sigma/\lambda < 0{,}1 \quad \text{et } \tau' = \sigma'/L < 0{,}1.$$

En faisant un développement limité à l'ordre 1 de x_i/y_i au voisinage de λ/L :

$$\frac{x_i}{y_i} = \frac{\lambda+\varepsilon_i}{L+\varepsilon'_i} \approx \frac{\lambda}{L}\times\left(1+\frac{\varepsilon_i}{\lambda}-\frac{\varepsilon'_i}{L}\right)$$

On approchera la variable X_i/Y_i par la variable aléatoire Z_i, avec

$$Z_i = \frac{\lambda}{L}\left(1+\frac{E_i}{\lambda}-\frac{E'_i}{L}\right) = \frac{\lambda}{L} + E''_i \text{ où } \quad E''_i = \frac{1}{L}\times\left(E_i - \frac{\lambda E'_i}{L}\right)$$

La variable Z_i suit une loi de Gauss d'espérance λ/L et de variance $\left(\frac{\lambda}{L}\right)^2\times\left(\tau^2+\tau'^2\right)$.

On notera que la loi exacte de X_i/Y_i n'est pas une loi de Gauss et que son espérance n'est pas exactement λ/L: on a approché sa loi par celle de Z_i, et ceci n'a de sens que parce que les nombres τ et τ' sont petits.

Si on ne dispose pas des 100 mesures, mais seulement des moyennes et écart-types empiriques des mesures de longueurs et largeurs, la quantité $s/\sqrt{n}$, avec ici n=100, pourra être calculée ainsi :

$$\frac{1}{\sqrt{100}}\times\frac{20{,}96}{29{,}71}\times\sqrt{\left(\frac{0{,}09}{20{,}96}\right)^2+\left(\frac{0{,}10}{29{,}71}\right)^2}$$

L'estimation de l'erreur standard est ici $3{,}8\times 10^{-4}$ et, pour un niveau de confiance 0,95, le résultat peut être écrit sous la forme :

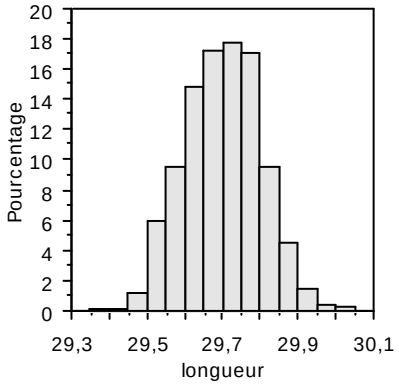
$$0{,}705 \pm 0{,}001$$

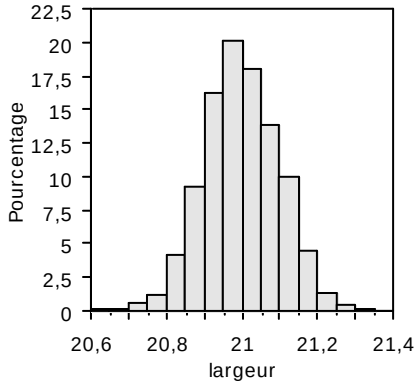
(Nous n'avons pas écrit ici $0{,}705 \pm 0{,}00038$: par souci de cohérence, le nombre de décimales de la moyenne et de l'erreur standard fournis dans les résultats seront les mêmes)

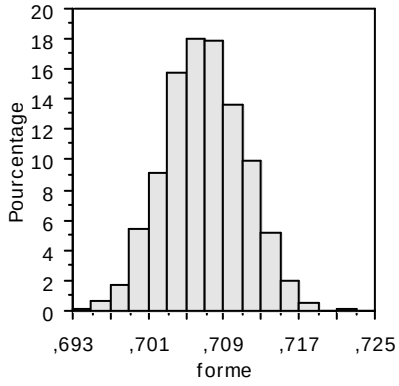
Si on dispose des n mesures, on peut évidemment calculer pour i=1…n, les nombres $z_i=x_i/y_i$. Ici, avec les données en jeu, la moyenne de $z_1,\ldots,z_{100}$ vaut 0,706 et l'erreur standard (quotient de l'écart-type par la racine du nombre de données) est 4×10^{-4}.

On passe du format Ai au format A$(i+1)$ en divisant la feuille Ai en deux feuilles d'aires égales (si λ_i et L_i sont la longueur et la largeur de Ai, alors la longueur de A$(i+1)$ est λ_i et sa largeur est $L_i/2$) ; de plus, le format (défini par le rapport largeur/longueur) est conservé par passage de Ai à A$(i+1)$ et la surface de A_0 est un mètre carré. On en déduit que le rapport donnant la forme est $1/\sqrt{2}\approx 0{,}707$ et que les dimensions de la feuille A4 sont, en cm, $100\times 2^{-1{,}75}\approx 21{,}02$ et $100\times 2^{-1{,}25}\approx 29{,}73$. Simulons des mesures de 1000 feuilles A4 a partir de lois de Gauss de moyennes 21,02 et 29,73 et d'écart-type 0,10 (le choix de l'écart-type pour cette simulation est ici proposé à partir des 100 mesures effectivement réalisées et décrites au début de ce paragraphe) ; les résultats sont résumés numériquement ci-dessous.

	Moy.	Dév. Std	Erreur Std	Nombre	Minimum	Maximum
largeur	20,999	,100	,003	1000	20,612	21,342
longueur	29,702	,100	,003	1000	29,381	30,042
forme	,707	,004	1,323E-4	1000	,693	,722







Si on ne dispose, pour la longueur et la largeur, que de la moyenne et de l'écart-type, l'erreur standard du rapport largeur /longueur peut être estimé par :

$$\frac{1}{\sqrt{1000}}\times\frac{21}{29,7}\times\sqrt{\left(\frac{0,1}{21}\right)^2+\left(\frac{0,1}{29,7}\right)^2}\approx 1,3\times 10^{-4}$$

On peut constater en se reportant au tableau ci-dessus, que cette valeur est égale, à la précision des calculs faits, à l'erreur standard calculée sur les 1000 rapports (largeur/longueur).

● Plus généralement, cherchons à estimer une grandeur δ liée à la grandeur μ par $\delta=f(\mu)$, où f est une fonction dérivable, dans un modèle où la loi de la mesure $X=\mu+E$ est gaussienne d'écart-type σ, avec σ/μ petit et E centrée. On approchera la loi de la variable $f(X)$ par celle de la variable $Z=f(\mu)+f'(\mu)E$, dont la loi est la même que celle de $T=\delta+|f'(\mu)|\times E$ (car $\delta=f(\mu)$ et si E suit une loi de Gauss centrée, il en est de même de $-E$) : l'habitude est de considérer T. La variable T suit une loi de Gauss d'espérance δ et d'écart-type $|f'(\mu)|\,\sigma$.

La grandeur δ peut aussi être fonction de plusieurs autres variables que l'on peut mesurer, $\delta=f(\mu_1,..,\mu_\rho)$, la fonction f étant différentiable, au voisinage de δ. Soient $X_1,\ldots, X_r$ les variables donnant les mesures de $\mu_1,..,\mu_\rho$ avec $X_i=\mu_i+\varepsilon_i$. On se place dans un modèle où les variables ε_i sont indépendantes, gaussiennes centrées, d'écarts-type σ_i, avec σ_i/μ_i petit. On approche la variable aléatoire $Z=f(X_1,\ldots, X_r)$ par la variable T, avec :

$$T=\delta+\left|\frac{\partial f}{\partial x_1}(\mu_1,..,\mu_r)\right|\times\varepsilon_1+...+\left|\frac{\partial f}{\partial x_r}(\mu_1,..,\mu_r)\right|\times\varepsilon_r$$

La loi de T est encore une loi de Gauss, d'espérance δ et de variance :

$$\sigma_T^2=\left|\frac{\partial f}{\partial x_1}(\mu_1,..,\mu_r)\right|^2\times\sigma_1^2+...+\left|\frac{\partial f}{\partial x_r}(\mu_1,..,\mu_r)\right|^2\times\sigma_r^2$$

Notons que si f est une forme linéaire, alors $T=Z$ et la formule ci-dessus n'est pas une approximation, mais donne la valeur exacte de la variance de Z.

Quand on estime les variances théoriques par les variances empiriques, on pourra donner, à partir des moyennes et variances empiriques sur chaque mesure, un résultat pour δ sous la forme :

$$f(\bar{x}_1,..,\bar{x}_r)\pm ks_T/\sqrt{n} \text{ avec } s_T^2=\left|\frac{\partial f}{\partial x_1}(\mu_1,..,\mu_r)\right|^2\times s_1^2+...+\left|\frac{\partial f}{\partial x_r}(\mu_1,..,\mu_r)\right|^2\times s_r^2$$

k valant 2 ou 3 suivant le niveau de confiance recherché.

En guise de conclusion, remarquons qu'en 1960, les méthodes de calculs probabilistes ci-dessus étaient déjà bien connues, mais sans doute *trop jeunes* pour être introduites dans l'enseignement de mathématiques ou de physique au lycée ; petit à petit, les calculs d'erreurs et d'incertitude ont disparu des programmes de physique : le caractère inéluctablement aléatoire de la mesure ne pouvait pas être abordé. Aujourd'hui, les élèves de lycée, s'ils ont peu l'occasion de faire un grand nombre de mesures de la même grandeur, rencontrent cependant la notion d'incertitude de mesure quand il s'agit par exemple de vérifier la loi d'Ohm $V=RI$: pour une intensité I_0 fixée, les points expérimentaux de coordonnées (r_k, v_k), $k=1..n$ ne sont pas exactement alignés et il convient alors d'en proposer une explication en considérant le caractère aléatoire de la mesure d'une grandeur.

Tests de différences en analyse sensorielle

Pascal Schlich
Centre Européen des Sciences du Goût, Dijon
Schlich@cesg.cnrs.fr

Introduction

L'analyse sensorielle consiste à décrire les propriétés organoleptiques (aspect, odeur, texture, saveur, arômes, …) et hédoniques (plaisir procuré au consommateur) des aliments et des boissons à l'aide de jurys de dégustation, plutôt appelés groupes ou panels d'évaluation sensorielle. L'un des objectifs de l'analyse sensorielle est d'établir si deux produits A et B sont perçus comme différents par un groupe d'évaluation sensorielle lors d'une dégustation à l'aveugle. Il s'agit de choisir et d'appliquer un test de différence, aussi appelé test de discrimination. Par exemple, l'épreuve triangulaire consiste à goûter 3 échantillons non distinguables visuellement dont on sait que deux sont le produit A alors que le troisième est le produit B; il s'agit pour le sujet de désigner l'échantillon provenant du produit non répété B. S'il ne perçoit pas de différence de goût entre ces trois échantillons, alors le sujet répondra au hasard et le nombre de bonnes réponses du panel sera alors distribué selon une loi binomiale de paramètre n (le nombre de sujets) et 1/3 (la probabilité de succès). Dans cette procédure, le sujet reçoit la consigne de toujours fournir une réponse, quitte à choisir un produit au hasard s'il pense ne percevoir vraiment aucune différence.

Une fois le test réalisé par les n sujets, on peut, à partir du nombre de bonnes réponses, calculer la probabilité de se tromper en déclarant les deux produits comme perçus différemment par les sujets (risque de première espèce, α). Mais l'analyse sensorielle utilise le plus souvent ce test afin de confirmer l'absence d'une différence perceptible entre les deux produits ; il convient dans ce cas de calculer une autre probabilité correspondant au risque de se tromper en déclarant les produits identiques. Le calcul de ce risque, appelé risque de seconde espèce β, nécessite de fixer a priori la grandeur d de la différence entre les produits que le test devrait être en mesure de détecter au niveau α du risque de première espèce. La probabilité $1-\beta$, appelée puissance du test, est donc celle de détecter une différence d entre les produits avec un risque inférieur à α de se tromper. Ainsi, la puissance apparaît comme une fonction de n, d et α

Après avoir introduit la loi binomiale dans le contexte de l'épreuve triangulaire, nous montrerons comment l'utiliser pour réaliser les calculs de risque de seconde espèce et de puissance. Nous présenterons alors des tables statistiques (Schlich, 1993) qui permettent à l'analyste sensoriel de bâtir ses expérimentations, c'est-à-dire de choisir le nombre de sujets à recruter en fonction des choix raisonnés des risques α et β.

Nous discuterons ensuite de l'intérêt de faire des répétitions, c'est-à-dire de demander à chaque sujet de réaliser plusieurs fois le même test. Nous évoquerons, sans rentrer dans les détails, comment le traitement statistique de ces données peut prendre en compte les répétitions.

Enfin, nous conclurons en montrant pourquoi le cadre de la loi binomiale n'est qu'un modèle peut-être trop simpliste face à la complexité de la situation d'un test de différence.

Définition et principe d'un test de différence

La loi de Bernoulli et la loi Binomiale

Dans l'épreuve triangulaire décrite en introduction, il existe deux alternatives, ou bien le sujet perçoit une différence entre les deux répétitions et on s'attend alors à ce qu'il donne la bonne réponse, ou bien il n'en perçoit pas et il donne alors une réponse au hasard qui a de fait une chance sur trois d'être la bonne. Une telle épreuve aléatoire à deux issues, échec ou succès (codées respectivement 0 ou 1), où la probabilité d'observer un succès est connue à l'avance et nommée p (ici p=1/3), est dite suivre une loi de Bernoulli de paramètre p. La probabilité d'observer un échec est alors égale à 1-p de sorte que la somme des probabilités des deux évènements possibles est bien égale à 1.

Lorsque l'on dispose d'une collection de n épreuves indépendantes chacune régie par la même loi de Bernoulli de paramètre p, la somme du nombre de succès suit par définition une loi Binomiale de paramètre n et p. Ainsi, le nombre de bonnes réponses fournies par un panel de n sujets à une même épreuve triangulaire est distribué selon une loi binomiale de paramètres n et 1/3, du moins si l'on fait l'hypothèse que les deux produits comparés dans le test ne sont pas perçus différemment par le groupe. Il convient d'insister sur l'importance de l'indépendance des épreuves de Bernoulli qui signifie pratiquement que chaque épreuve ne doit pas être influencée par le résultat de chacune des autres épreuves ; en clair, il suffit de s'assurer que les sujets de l'épreuve triangulaire ne communiquent pas entre eux au cours de la dégustation.

En notant X_i le résultat (codé 0 ou 1) de la i-ème épreuve de Bernoulli, le nombre total de succès, qui suit la loi Binomiale, est noté: $X=\sum_{i=1}^{n} X_i$ et est une variable aléatoire discrète prenant ses valeurs entre 0 et n. Grâce à l'indépendance des épreuves, la probabilité de chacune de ces valeurs est donnée par :

$$P(X=x)=\binom{n}{x}p^x(1-p)^{n-x} \qquad (1)$$

où $\binom{n}{x}=\frac{n!}{x!(n-x)!}$ est le nombre de façons de choisir les x épreuves à succès parmi les n épreuves.

Le principe du test statistique

Si réellement les deux produits ne sont pas distinguables au goût, alors le nombre de bonnes réponses du panel est distribué selon cette loi binomiale de paramètre n et 1/3. Cette situation est appelée l'hypothèse H_0, dite " hypothèse nulle ". On s'attend à observer " en moyenne " un nombre de bonnes réponses proche de n/3. Cette valeur moyenne attendue est désignée en statistique par le terme d'espérance mathématique de la variable aléatoire X. On peut calculer l'espérance mathématique de toute variable aléatoire discrète en additionnant toutes les réalisations possibles de la variable aléatoire chacune étant multipliée (c'est-à-dire pondérée) par sa probabilité de réalisation.

Si au contraire les sujets, ou du moins certains d'entre eux, sont capables de faire la différence entre les deux produits, alors le nombre de bonnes réponses du panel est attendu supérieur à n/3. Le but du test statistique est de décider à partir de quel nombre de bonnes réponses observées x on peut considérer que les produits sont perçus différemment par le panel en tant qu'unité. Pour cela, on considère la probabilité suivante :

$$P(X \geq x) = \sum_{i=x}^{n} P(X = i) \qquad \textbf{(2)}$$

qui est celle d'observer au moins x bonnes réponses. Cette probabilité est d'autant plus petite que x est grand. Lorsque qu'elle devient suffisamment petite, cela signifie que le nombre de bonnes réponses est devenu suffisamment grand pour remettre en cause l'hypothèse nulle. Il s'agit là du principe de base des tests statistiques : **on calcule une statistique, dont on connaît la distribution de probabilité sous l'hypothèse nulle, afin de pouvoir infirmer l'hypothèse nulle si la valeur observée de la statistique devient peu probable.**

Classiquement, on demande à la probabilité donnée par la formule (2) d'être inférieur à 0.05 pour rejeter l'hypothèse nulle et donc pour déclarer les deux produits différents ; le risque de se tromper, aussi appelé risque de première espèce α ou niveau du test, en prenant cette décision est alors inférieure à 5%. On peut aussi décider de faire un test au niveau 10%, voire 20%, selon le niveau du risque que l'on est prêt à prendre en déclarant les produits différents alors qu'en fait ils ne l'étaient pas. Il faut bien comprendre qu'en général l'hypothèse nulle devrait être l'hypothèse dangereuse, celle dont le préjudice d'un rejet à tort aurait un coût dont l'importance va nous guider dans le choix d'un risque de première espèce. Par exemple, l'hypothèse nulle " le patient est malade ", car si l'on conclut qu'il ne l'est pas, il ne s'agit pas de se tromper ; ou encore, " la sécurité de cette centrale nucléaire est déficiente ", hypothèse que l'on souhaitera même tester au risque, non pas de 5%, mais de 1%, voire 1‰. En revenant à l'épreuve triangulaire, rejeter l'hypothèse nulle, c'est rejeter que les produits ne soient pas distinguables et c'est donc prendre le risque de refuser de substituer le produit B au produit A, alors que son coût de revient est peut être bien inférieur.

Selon cette logique, l'épreuve triangulaire devrait être mise en œuvre dans des situations où l'on pense que la différence existe et où il est important de le démontrer. Par exemple, mon produit perd des parts de marché depuis l'arrivée d'un nouveau produit concurrent sur le marché. Après avoir dégusté le nouveau produit avec mon équipe, je pense qu'il a effectivement un goût différent du nôtre et que c'est peut-être la raison de notre perte de parts de marché. Toutefois, avant de me lancer dans la modification de la formule de notre produit ; je souhaiterais m'assurer que la population de consommateurs de ce type de produits fait bien la différence entre les deux produits en concurrence. Je vais donc organiser un test de différence en choisissant un risque de première espèce petit, par exemple 5%, afin de n'engager un coûteux programme de recherche visant à reproduire le nouveau produit, que si je suis quasiment certain que les deux produits étaient vraiment perçus différemment par les consommateurs.

Lorsque j'aurai mis au point notre nouveau produit, je vais être confronté au problème inverse du précédent, à savoir démontrer la similitude gustative de ce produit avec celui de nos concurrents. Je vais alors à nouveau organiser un test de différence, mais cette fois-ci dans le but de ne pas rejeter l'hypothèse nulle et surtout en contrôlant le risque β, dit risque de seconde espèce, de me tromper en déclarant les produits identiques alors qu'en fait ils ne l'étaient pas. Il s'agit là du problème du contrôle de la puissance d'un test statistique

La puissance du test

Définition

On appelle puissance d'un test sa probabilité de détecter une différence existante. C'est donc la probabilité de rejeter à juste titre l'hypothèse nulle, soit la probabilité complémentaire du risque β d'accepter à tort cette hypothèse. Ainsi la puissance d'un test est égale à 1-β.

Si l'on désigne par H_1, l'hypothèse alternative à l'hypothèse nulle H_0, le tableau 1 résume les différentes situations possibles :

Tableau 1. Les risques en jeu lors d'un test statistique

		Décision	
		H_0	**H_1**
Réalité	**H_0**	OK	Risque de première espèce α
	H_1	Risque de seconde espèce β	Puissance 1-β

Pour calculer la puissance, il faut spécifier H_1 c'est-à-dire la hauteur de la différence à mettre en évidence. Dans le cadre de l'épreuve triangulaire, il s'agit de fixer une probabilité p_1 de réussite supérieure à la probabilité p=1/3 de réussite sous le seul effet du hasard, c'est-à-dire en l'absence de différence perceptible entre les produits. Les hypothèses H_0 et H_1 du test statistique associé à l'épreuve triangulaire sont donc :

H_0 : p=1/3 H_1 : $p=p_1=1/3+d$ avec d>0 représente la différence entre les deux produits

En se fixant un risque α, on a vu que l'on pouvait déterminer en utilisant les formules (1) et (2) le nombre x de bonnes réponses à obtenir pour rejeter H_0. Ce nombre x étant fixé, la donnée de p_1 permet maintenant de calculer β de la manière suivante :

$$\beta = P(X < x \quad | \quad p = p_1) = \sum_{i=0}^{x-1} \binom{n}{i} p_1^i (1-p_1)^{n-i} \qquad (3)$$

C'est-à-dire en sommant les probabilités d'observer 0, 1, 2, …, x-1 bonnes réponses, soit l'ensemble des évènements qui conduisent à ne pas rejeter H_0, alors qu'en fait c'est H_1 qui est vraie, justifiant l'emploi de p_1 à la place de p dans la formule binomiale.

Tables statistiques pour le contrôle des risques α et β

Les tables statistiques à la disposition des praticiens de l'analyse sensorielle ne présentaient classiquement que le nombre de réponses correctes ***x*** parmi ***n*** à obtenir pour pouvoir rejeter l'hypothèse nulle avec un risque ***α*** inférieur à 10%, 5% ou 1%. Nous avons proposé (Schlich, 1993) deux séries de tables permettant de contrôler simultanément les risques ***α*** et ***β*** dans le

cadre de l'épreuve triangulaire (p=1/3) mais aussi pour d'autres épreuves de différences correspondantes à p=1/2. Pour cela, nous avons choisi d'exprimer d, la différence entre les deux produits à mettre en évidence, sous la forme de p_c appelée la proportion corrigée du hasard des bonnes réponses et donnée par la formule suivante :

$$p_c = (p_1 - 1/3)/(2/3) \qquad (4)$$

En remplaçant p_1 par la proportion observée de bonnes réponses, on comprend p_c comme la proportion de sujets qui font réellement (sans l'aide du hasard) une différence entre les produits. Dans la première série de tables, dont le tableau 2 présente un extrait, nous avons fixé 3 valeurs pour p_c qui, exprimées en pourcentage, étaient de 50%, 37,5% et 25% et que nous avons arbitrairement appelées grande, moyenne et petite différences. Pour n variant de 5 à 50, nous avons alors tabulé les risques α et β associés à chaque règle de décision définie par le rejet de l'hypothèse nulle si x bonnes réponses sont observées. Nous n'avons pas inclus les x conduisant à α supérieur à 50% et nous n'avons pas imprimé de β supérieur à 50%.

Tableau 2. Extrait de l'annexe 1 de Schlich (1993)

APPENDIX 1. Table of risks (%) for triangular test.

			β		
n	x	α	50%	37·5%	25%
5	3	21	21	35	50
6	3	32	10	20	34
	4	10	32	49	
7	3	43	5	11	23
	4	17	17	32	50
	5	5	43		
8	4	26	9	20	36
	5	9	26	44	
9	4	35	4	12	25
	5	14	14	30	
	6	4	35		
10	4	44	2	7	17
	5	21	8	20	38
	6	8	21	41	
	7	2	44		
11	5	29	4	12	27
	6	12	12	28	50
	7	4	29		
21	8	40	0	2	9
	9	24	1	5	19
	10	12	2	11	33
	11	6	6	22	50
	12	2	12	37	
	13	1	24		
	14	0	40		
22	8	46	0	1	7
	9	29	0	3	14
	10	16	1	8	26
	11	8	3	16	42
	12	3	8	28	
	13	1	16	44	
	14	0	29		
	15	0	46		
23	9	35	0	2	11
	10	21	1	5	20
	11	11	2	11	34
	12	5	5	21	50
	13	2	11	35	
	14	1	21		
	15	0	35		
30	11	42	0	1	5
	12	28	0	1	10
	13	17	0	3	18
	14	9	1	7	29
	15	4	2	13	43
	16	2	4	23	
	17	1	9	35	
	18	0	17	50	
	19	0	28		
	20	0	42		
31	11	47	0	0	4
	12	32	0	1	7
	13	20	0	2	14
	14	12	0	5	24
	15	6	1	10	36
	16	3	3	17	
	17	1	6	28	
	18	0	12	41	
	19	0	20		
	20	0	32		
	21	0	47		
32	12	37	0	1	6
	13	24	0	1	11

Les 3 dernières colonnes de ces tables donnent le risque β associé successivement aux valeurs de p_c de 50 %, 37,5 % et 25 %. Par exemple avec 30 sujets, un test réalisé au niveau α de 5 % correspondrait à une règle de décision de x=15 bonnes réponses, mais ce test n'aurait qu'une probabilité de 57 % (risque β de 43 %) d'être significatif si 25 % des sujets faisaient réellement une différence entre les produits, alors qu'il détecterait presque à coup sur (β=2 %) la différence si un sujet sur deux (p_c=50 %) percevait réellement la différence. On voit bien ici l'importance du paramètre p_c. Si le choix de n variant de 5 à 50 correspond bien à la réalité des tailles de panels d'analyse sensorielle, celui des 3 valeurs de p_c ne correspond pas toujours à la réalité économique. Il suffit parfois de 5 % de consommateurs qui ne retrouve plus le goût de leur produit usuel et donc qui menacent de changer de marque pour fragiliser la rentabilité de cette marque.

L'intérêt de cette première série de table est également pédagogique car la tabulation simultanée du risque α associé à chaque nombre de bonnes réponses avec 3 risques β correspondant à 3 hypothèses H_1 différentes est utile lorsqu'il s'agit d'expliquer le concept de test statistique.

Afin de couvrir un éventail plus large de valeurs de p_c, une seconde série de tables a été proposée dans le même article. Le tableau 3 donne cette table pour le test triangulaire. On peut se rendre compte que s'il s'agit de détecter un p_c de 5 % avec une probabilité 1-β de 95 %, alors même en tolérant un risque élevé α de 20 % il faudra rassembler un panel d'au moins 1273 sujets ! En conservant les mêmes risques α et β, il est intéressant de regarder l'évolution de n: 1273, 325, 86, 39, 25 et 16 selon que p_c augmente de 5 % à 50 %. Ceci illustre bien que la taille de la différence à mettre en évidence est vraiment le concept clef pour le contrôle de la puissance. Ces tables ont eu le mérite d'alerter des analystes sensoriels sur l'inadéquation à leur problème pratique des tests de différence qu'ils utilisaient en routine.

Tableau 3. Annexe 3 de Schlich (1993)

APPENDIX 3. Triangular test. Minimum total number of responses (n) and associated number of correct responses (x) for a range of type 1 and type 2 risks.

Risk 1	Risk 2 (1-power)									
	0·200		0·100		0·050		0·010		0·001	
	n	x	n	x	n	x	n	x	n	x
					p_c = 50%					
0·200	7	4	12	6	16	8	25	11	36	15
0·100	12	7	15	8	20	10	30	14	43	19
0·050	16	9	20	11	23	12	35	17	48	22
0·010	25	15	30	17	35	19	47	24	62	30
0·001	36	22	43	25	48	27	62	33	81	41
					p_c = 40%					
0·200	12	6	17	8	25	11	36	15	55	22
0·100	17	9	25	12	30	14	46	20	67	28
0·050	23	12	30	15	40	19	57	26	79	34
0·010	35	19	47	24	56	28	76	36	102	46
0·001	55	30	68	36	76	39	102	50	130	61
					p_c = 30%					
0·200	20	9	28	12	39	16	64	25	97	37
0·100	30	14	43	19	54	23	81	33	119	47
0·050	40	19	53	24	66	29	98	41	136	55
0·010	62	30	82	38	97	44	131	57	181	76
0·001	93	46	120	57	138	64	181	81	233	101
					p_c = 20%					
0·200	39	16	64	25	86	33	140	52	212	77
0·100	62	26	89	36	119	47	178	68	260	97
0·050	87	37	117	48	147	59	213	83	305	116
0·010	136	59	176	74	211	87	292	117	397	155
0·001	207	91	257	110	302	127	396	162	513	205
					p_c = 10%					
0·200	149	55	238	86	325	116	529	186	819	285
0·100	240	90	348	128	457	166	683	244	1011	357
0·050	325	123	447	166	572	210	828	299	1181	421
0·010	525	201	680	256	824	307	1132	415	1539	557
0·001	803	310	996	379	1165	439	1530	568	1992	730
					p_c = 5%					
0·200	593	208	931	323	1273	439	2074	710	3201	1090
0·100	927	328	1359	476	1761	613	2682	926	3943	1353
0·050	1271	452	1763	621	2225	779	3246	1127	4616	1592
0·010	2059	737	2660	944	3236	1142	4440	1554	6021	2093
0·001	3162	1137	3903	1393	4585	1628	6006	2116	7827	2739

p_c, Percentage above chance.

Compenser le manque de sujets par des répétitions

Lorsque le nombre de sujets requis dépasse les moyens disponibles en pratique, la tentation est forte de compenser le manque de sujets par des répétitions, c'est-à-dire de donner plusieurs fois le même test à chaque sujet. Si par exemple n sujets ont effectué k répétitions chacun du même test est-il correct de traiter cette situation comme si $n.k$ sujets avaient réalisé chacun une seule fois ce test ? La réponse est bien évidemment négative. Cependant, il n'existe pas encore de consensus parmi les théoriciens de l'analyse sensorielle sur la manière de traiter ce problème. Nous allons brièvement décrire l'état de la controverse.

Il convient tout d'abord de remarquer que sous l'hypothèse nulle chaque sujet a bien une même probabilité de 1/3 de donner une bonne réponse à chaque répétition, si l'on suppose que les répétitions d'un même sujet sont indépendantes entre elles, alors le nombre total de bonnes réponses suit bien une loi binomiale de paramètres $n.k$ et 1/3 et il est fort tentant de tester H_0 en utilisant cette loi, on appellera ce test le " test optimiste ". Pourtant, nous pensons que ce test n'est pas correct dans la mesure où nous n'avons plus à faire à un échantillon aléatoire d'une population cible, mais à un échantillon à deux strates (les sujets et les répétitions). Poussé à l'extrême, le test optimiste pourrait conduire à n'utiliser qu'un seul sujet effectuant $\boldsymbol{n.k}$ répétitions…. Ce qui serait évidemment grotesque.

A l'opposé, on peut penser à un test " pessimiste " qui consisterait à ne retenir pour chaque sujet que sa proportion de réussite sur l'ensemble des répétitions et d'utiliser la somme de ces proportions individuelles comme nombre de réussite dans un test binomial classique de pamètres ($\boldsymbol{n}$, 1/3). Mais sous l'hypothèse H_1 où les deux produits sont perçus comme différents, le test pessimiste, comme le test optimiste d'ailleurs, ne resterait valable que sous la condition que les probabilités de réussite de chacun des sujets soient toutes égales, on parle alors d'homogénéité de la population quant à sa capacité à détecter la différence. Ainsi, les calculs de puissance, qui nécessitent de spécifier une hypothèse alternative H1, ne seraient valables que pour une population homogène quant à sa probabilité de détecter la différence.

Brockhoff et Schlich (1998) sont partis du principe que le test optimiste était forcément la stratégie la plus puissante possible, alors que le test pessimiste est la moins puissante. Puisque l'hétérogénéité de la population quant à la probabilité de réussir le test constitue le frein principal à l'utilisation d'une stratégie binomiale permettant des calculs de puissance, ces auteurs ont proposé un test qui coïncide avec le test optimiste pour une population parfaitement homogène et avec le test pessimiste pour une population complètement hétérogène. Ce test consiste à réaliser un test binomial classique en considérant que le nombre total de réponses était égal à $\boldsymbol{n.k/\sigma^2}$, ou à son nombre entier le plus proche, et que le nombre de bonnes réponses était lui égal à $\boldsymbol{x/\sigma^2}$; où σ^2 est un coefficient de *sur-dispersion* qui consiste à estimer le facteur multiplicatif de la variance de la loi binomiale occasionné par le fait que chaque sujet n'a pas eu le même taux de réussite aux $\boldsymbol{k}$ répétitions. Ce coefficient d'hétérogénéité est compris entre 1 pour une population parfaitement homogène et $\boldsymbol{k}$ pour une population complètement hétérogène.

Le test de Brockhoff et Schlich requiert une estimation du coefficient σ^2 selon des formules fournies par les auteurs, mais qui nécessitent que des données issues d'un test avec répétitions soient déjà disponibles. Or dans la phase de planification du test, l'analyste sensoriel doit contrôler la puissance de son test par un choix raisonné du nombre de sujets et de répétitions qui nécessitera une estimation a priori de ce coefficient. Ce problème est débattu par Schlich *et al.* (2000) sur la base de 6 jeux de données de tests triangulaires intensément répétés (entre 8 et 60 répétitions) avec des groupes de sujets composés de 12 à 61 étudiants. Il ressort de ces travaux qu'une estimation correcte du véritable niveau d'hétérogénéité d'une population requiert une première expérimentation basée sur un grand nombre de répétitions (environ une dizaine), ce qui est rarement faisable.

Le test de Brockhoff et Schlich a été discuté et critiqué par Kunert et Meyners (1999). Il n'est guère possible de rentrer dans le détail de cette discussion ici, mais il faut savoir que celle-ci n'est pas close, puisque Brockhoff va prochainement publier un nouvel article sur ce sujet qui devrait relancer le débat de manière contradictoire entre statisticiens et praticiens de l'analyse sensorielle. D'autre part, Ennis et Bi (1998) proposent une autre approche basée sur la loi beta-binomiale, qui correspond assez bien à la réalité d'un test de différence avec répétitions.

Enfin, il convient d'ajouter qu'aucun de ces tests ne prend en compte que les résultats successifs d'un même sujet à un même test au fil d'une même séance sont peut-être auto-corrélés en raison de phénomène d'apprentissage ou au contraire de fatigue sensorielle.

Cette discussion montre qu'il serait prématuré de vouloir ériger aujourd'hui l'une ou l'autre de ces méthodes en une norme AFNOR ou ISO, même si les praticiens de l'analyse sensorielle réclament cette norme avec insistance. Elle montre aussi qu'un paradigme aussi simple que l'épreuve triangulaire soulève en fait une foule de questions se situant à la frontière des statistiques, de la psychophysique et de la psychologie.

Références :

Brockhoff P. B. et Schlich P. (1998). Handling replications in discrimination tests. *Food Quality and Preference*, **9,** 303-312.

Ennis D.M. et Bi J. (1998). The beta-binomial model : accounting for inter-trial variation in replicated difference and preference tests. *Journal of Sensory Studies*, **13**, 389-412.

Kunert J. et Meyners M. (1999). On the triangle test with replications. *Food Quality and Preference*, **10**, 477-482.

Schlich P. (1993). Risk tables for discrimination tests. *Food Quality and Preference*, **4**, 141-151.

Schlich P., Dacremont C. et Brockhoff P.B. (2000). Application of replicated difference testing. *Food Quality and Preference*, **11**, 43-46

LA NOTATION STATISTIQUE DES EMPRUNTEURS OU " SCORING "

Gilbert Saporta
Conservatoire National des Arts et Métiers
saporta@cnam.fr

Dans leur quasi totalité, les banques et organismes financiers utilisent l'analyse statistique pour prédire si un emprunteur sera un bon ou un mauvais payeur et prendre ensuite la décision appropriée : acceptation sans condition, prise de garantie, refus.

La modélisation et la décision se fondent sur l'observation du passé : on connaît pour un certain nombre de prêts attribués la qualité payeur qui est donc une variable qualitative Y à deux modalités (" bon " ou " mauvais ") ainsi que les données recueillies lors du dépôt du dossier de prêt : ce sont les variables X (X_1, …, X_p) . Typiquement pour des particuliers on trouvera l'age, la profession , le statut matrimonial, le fait d'être ou non propriétaire, donc majoritairement des variables qualitatives, alors que pour des entreprises on aura plutôt des variables numériques comme des ratios issus de la comptabilité.

Formellement il s'agit de trouver une fonction f(X_1, …, X_p) permettant de prédire Y.

Dans ce qui suit nous décrirons les diverses étapes et les problèmes qui se posent depuis la collecte des données jusqu'à la mise en œuvre en donnant à chaque fois des indications sur les méthodologies à utiliser.

I. La collecte de l'information

Le premier travail consiste à constituer un fichier qui contient des informations complètes sur des dossiers de prêts. Il se présentera sous la forme d'un tableau rectangulaire individus-variables où les n individus sont partagés en deux groupes d'effectifs n_1 et n_2 : les bons et les mauvais.

Ce travail essentiel est maintenant facilité par le stockage informatique, mais cela n'a pas toujours été le cas : les variables du dossier de demande n'étaient pas forcément saisies car elles n'étaient pas toutes jugées utiles pour la gestion du prêt. Il fallait alors retrouver les dossiers papiers.

Les n individus constituent en fait un échantillon de l'ensemble des N données disponibles : nous verrons plus loin qu'il est indispensable de garder de côté un certain nombre de dossiers afin de valider les résultats obtenus. Il faut donc prélever aléatoirement n individus parmi les N : comme il faut s'assurer d'avoir un nombre suffisant et non aléatoire (ce qui introduirait une source de variabilité supplémentaire, donc une moindre précision) d'observations dans chacun des deux groupes, on procède à un sondage stratifié avec tirage séparé des n_1 et n_2 individus. Deux questions se posent alors : quel effectif global et quelle répartition de n_1 et n_2 ? Une idée naturelle consisterait à prélever n_1 et n_2 en respectant les proportions de bons et mauvais dossiers, d'autant plus que l'on sait que le sondage stratifié à répartition proportionnelle est toujours meilleur que l'échantillonnage simple sans stratification. Cette méthode est cependant à déconseiller ici car les deux groupes ont des proportions très différentes : le groupe à risque (les mauvais payeurs) qu'il faut détecter est très minoritaire (mettons 10%) et serait mal représenté. On a pu démontrer qu'une répartition équilibrée $n_1 = n_2$ est bien meilleure, sinon optimale sous des hypothèses assez générales. Les vraies proportions p_1 et p_2 servent ultérieurement pour les calculs de probabilités *a posteriori*.

Quant au nombre total n, il est typiquement de quelques milliers.

Un problème plus complexe est celui du biais de sélection : en fait les dossiers dont on connaît l'issue (bons ou mauvais) résultent d'un choix effectué en général par des analystes de crédit ; tous les dossiers de prêt n'étaient évidemment pas acceptés et ceux qui l'ont été ne constituent pas un échantillon représentatif de toutes les demandes. Même si la méthode antérieure de sélection n'était pas scientifique, il est clair que les dossiers acceptés n'ont pas les mêmes caractéristiques que les dossiers refusés. Or pour construire une règle de décision valable pour tous les nouveaux dossiers, il aurait fallu savoir ce que seraient devenus les dossiers refusés si on les avait acceptés…Il faut alors recourir à des techniques assez élaborées (estimation en deux phases, modèle Tobit). Sans entrer dans les détails, disons seulement que l'on modélise également le processus de sélection.

Le problème du biais de sélection n'intervient pas dans d'autres domaines où des techniques similaires de scoring sont utilisées comme l'assurance automobile (pour la détection des conducteurs à risque) ou la sélection d'adresses pour optimiser l'envoi de propositions commerciales (dans ce dernier cas on effectue un scoring à partir des résultats d'un premier courrier ; les " bons " étant les répondants, les " mauvais " les non-répondants) .

II Les analyses préliminaires

Le fichier brut une fois constitué doit d'abord être " nettoyé " pour éliminer erreurs et incohérences. Il comporte alors en général un trop grand nombre de variables. Une exploration des liaisons entre chaque variable X et le critère à prédire Y permet en général d'éliminer les variables non pertinentes. On utilise alors des outils classiques : test du khi-carré de liaison entre variables qualitatives, comparaison des % de bons et de mauvais par catégorie de chaque variable X.

Dans le même temps on procède à des recodages des variables : regroupement de valeurs en classes pour les variables continues (on s'aide d'histogrammes), regroupement de classes pour obtenir la meilleure séparation sur Y. On crée également de nouvelles variables par combinaison de 2 ou plusieurs variables. Par exemple si on s'aperçoit que l'ancienneté dans l'emploi joue différemment selon la profession sur la probabilité de bon remboursement, on créera une variable croisant les modalités de ces deux variables (cf. exemple plus loin).

Il est couramment admis que toutes ces analyses représentent près de 80% du temps de ce genre d'études.

III La modélisation

Les techniques de " scoring " qui sont les plus utilisées dans le secteur bancaire utilisent des méthodes linéaires pour leur simplicité et leur grande robustesse. Il existe bien d'autres méthodes non-linéaires ou non-parametriques comme les arbres de décision, les réseaux neuronaux etc. dont l'usage se répand (cf. références) mais elles sortent de ce bref exposé.

Un score est une note de risque que l'on calcule comme combinaison linéaire des variables explicatives $S = a_0 + \sum_{i=1}^{p} a_i X_i$. Les coefficients a_i sont optimisés pour la prédiction de Y. Notons qu'ici le terme constant a_0 peut être omis.

Pour obtenir le vecteur **a** des coefficients des a_i , il existe diverses techniques d'estimation dont les deux principales sont la fonction linéaire discriminante de Fisher et le modèle logit (encore appelé régression logistique).

III.1 La fonction linéaire discriminante de Fisher.

C'est la plus ancienne (elle remonte à 1936) : c'est la combinaison optimale qui sépare le mieux les moyennes du score dans les deux groupes. Plus précisément si $\bar{s}_1$ et $\bar{s}_2$ sont les scores moyens sur les deux groupes de n_1 et n_2 individus, on maximise $\frac{(\bar{s}_1 - \bar{s}_2)^2}{V(s)}$ où *V(s)* est la moyenne pondérée des variances du score dans chacun des 2 groupes. On montre que **a** est proportionnel à $W^{-1}(g_1-g_2)$ où W est la moyenne pondérée des matrices de variance-covariance des variables explicatives dans chaque groupe et les g les vecteurs des moyennes des variables de chaque groupe. C'est une méthode de moindres carrés.

III.2 La régression logistique ou modèle logit .

On exprime la probabilité *a posteriori* d'appartenance à un des groupes selon :

$$P(G_1 / X) = \frac{\exp(S)}{1+\exp(S)} = \frac{\exp(a_0 + \sum_{i=1}^{p} a_i X_i)}{1+\exp(a_0 + \sum_{i=1}^{p} a_i X_i)}$$

et on estime alors les a_i par la méthode du maximum de vraisemblance. X désigne ici le vecteur dont les composantes sont les X_i pour i=1 à p.

Nous avons employé le terme de probabilité *a posteriori* qui renvoie à l'usage de la formule de Bayes. En effet si on connaît les probabilités *a priori* d'appartenance aux deux groupes p_1 et p_2=1- p_1, qui sont en fait les proportions réelles des deux groupes, la probabilité d'appartenir au groupe 1 connaissant les informations fournies par le dossier, c'est à dire les X, est donnée par :

$P(G_1 / X = x) = \frac{p_1 f_1(x)}{p_1 f_1(x) + p_2 f_2(x)}$ où f_k est la densité de probabilité des X dans le groupe k.

Pour de nombreux modèles probabilistes (gaussiens, multinomial etc.) cette probabilité *a posteriori* se met sous la forme logistique précédente : $P(G_1 / X) = \frac{\exp(S)}{1+\exp(S)}$

En particulier si le vecteur aléatoire des X suit une loi normale de même matrice de variance-covariance dans les deux groupes, la règle qui consiste à classer une observation x dans le groupe qui a la plus forte probabilité *a posteriori* est équivalente à la règle qui consiste à classer une observation dans un groupe selon que son score est inférieur ou supérieur à un certain seuil.

Les deux méthodes.(Fisher et logit) ne conduisent pas aux mêmes estimations des coefficients, mais celles-ci sont en général assez proches. Le choix entre les deux ne doit pas être une question d'école : moindres carrés contre maximum de vraisemblance, mais plutôt se faire sur leur capacité prédictive, c'est à dire sur de nouvelles observations.

La règle " naïve " de Bayes qui consiste à prédire le groupe le plus probable, donc ici à choisir le groupe qui a une probabilité *a posteriori* supérieure à 0.5, n'est en général pas adaptée à la prédiction d'un groupe rare. On cherche plutôt à détecter un maximum d'individus à risque, et on choisira le seuil de décision en conséquence (voir plus loin).

III.3 Cas de prédicteurs qualitatifs.

Le cas où les variables explicatives X_i sont qualitatives nécessite un traitement particulier. En effet comment faire une combinaison linéaire de variables qualitatives ? Cela n'a évidemment pas de sens. La solution retenue est basée sur ce que l'on appelle la forme disjonctive d'une variable qualitative X à m modalités (comme une profession). On définit les m variables indicatrices des modalités ($\mathbf{1}_1$, $\mathbf{1}_2$, ….,$\mathbf{1}_m$) telles que $\mathbf{1}_j$ vaut 1 si on appartient à la modalité j, 0 sinon. Seule une des indicatrices vaut 1, celle qui correspond à la modalité prise. Les *m* indicatrices sont donc équivalentes à la variable qualitative. Le score est alors une combinaison linéaire des indicatrices, ce qui revient à donner une note partielle à chaque modalité de chaque variable. Le score final étant la somme des notes partielles (à telle profession correspond telle note). Les variables explicatives qui interviennent dans les formules sont donc les indicatrices de toutes les variables.
Une difficulté intervient cependant : la matrice *W* n'est pas de plein rang et n'est donc pas inversible car la somme des indicatrices des modalités de chaque variable vaut 1. Cela signifie qu'il existe une infinité de solutions équivalentes pour estimer les coefficients :une des solutions couramment utilisée consiste alors à ne prendre que *m-1* indicatrices pour chaque variable qualitative puisque la dernière est redondante.

III.4 Un exemple

Les valeurs suivantes sont fictives (mais réalistes) et ne servent qu'à illustrer la méthode. Considérons le cas d'un établissement financier qui veut prédire la solvabilité d'entreprises pour savoir s'il doit ou non accorder un prêt. On connaît pour chaque entreprise les deux variables suivantes : X_1 part des frais financiers dans le résultat en %, et X_2 délai de crédit fournisseurs (nombre de jours avant de payer les fournisseurs) .

Sur l'échantillon des entreprises solvables la moyenne de X_1 vaut 40, celle de X_2 90. Sur l'échantillon des entreprises non solvables ces moyennes sont respectivement 90 et 100. On admet que les écart-types sont les mêmes d'un groupe à l'autre et sont respectivement s_1=40, s_2=20, et que X_1 et X_2 présentent la même corrélation r=0.8 dans chaque groupe. La covariance entre X_1 et X_2 vaut rs_1s_2=640.

La matrice de variance commune (dite également intra-classe) est alors $W = \begin{pmatrix} 1600 & 640 \\ 640 & 400 \end{pmatrix}$

et le vecteur de différence des moyennes $g_1 - g_2 = \begin{pmatrix} -50 \\ -10 \end{pmatrix}$

Il est facile d'en déduire la fonction de Fisher par la formule $a = W^{-1}(g_1-g_2)$. Les coefficients étant définis à une constante multiplicative près, on peut prendre pour a le vecteur de composantes –1 et 1.2 .

La fonction de score est alors $S = -X_1 + 1.2\ X_2$

On en déduit facilement par transformation linéaire que le score moyen des entreprises solvables vaut 68 tandis que le score moyen des entreprises non solvables vaut 30. Les écart-types des variables étant supposés identiques dans les deux groupes on trouve que $V(S) = V(X_1) + 1.2^2 V(X_2) - 2(1.2)\,\mathrm{cov}(X_1; X_2) = (25.3)^2$

On supposera pour la simplicité de l'exposé que la distribution du score suit dans chaque groupe une loi normale. Quand il n'en est pas ainsi, les densités de probabilité, les fonctions de répartition, etc. doivent être estimées d'une autre manière.

Un usage classique dans les études de ce type est de recaler le score S pour qu'il prenne la quasi totalité de ses valeurs dans l'intervalle [0 ; 1000]. Cela se fait simplement par transformation affine.

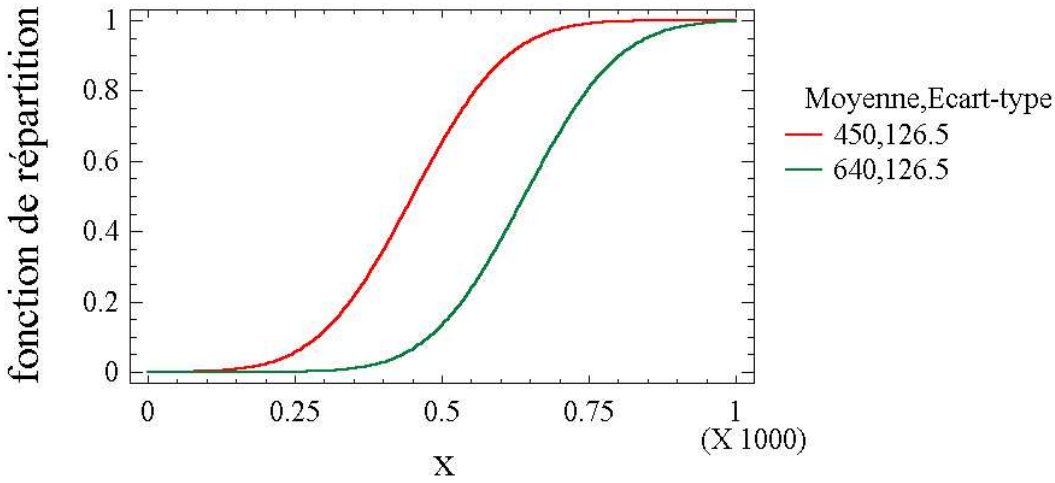
Ceci peut être réalisé approximativement dans notre exemple en multipliant le score par 5 et en ajoutant 300.

La fonction de score vaut donc $S= -5X_1 + 6 X_2 + 300$

V. Qualité et utilisation d'un score

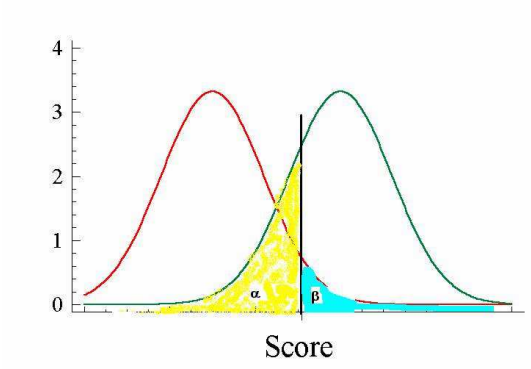
On estime tout d'abord les distributions conditionnelles du score dans chacun des deux groupes. Un score efficace doit conduire à des distributions bien séparées. Dans l'exemple précédent le score suit une loi normale N(640 ;126.5) pour le groupe des entreprises solvables ou une loi N(450 ;126.5) pour les entreprises non-solvables. On vérifiera que le score *S* donne une meilleure séparation que chaque variable prise séparément en calculant l'écart réduit entre moyennes, c'est à dire la différence en valeur absolue entre moyennes divisée par l'écart-type commun.

On considérera également les fonctions de répartition :



L'utilisation est la suivante : si on refusait de prêter de l'argent aux entreprises ayant une note de score inférieure à 556, on éliminerait 80% des entreprises insolvables (les " mauvaises ") mais on refuserait à tort 25% des entreprises solvables (les " bonnes "). Le choix du seuil dépend des risques financiers et est fixé par un raisonnement économique prenant en compte les coûts d'erreur de mauvaise classification : en effet accorder un prêt à une entreprise qui se révélera insolvable a un coût différent de celui de perdre un bon client.

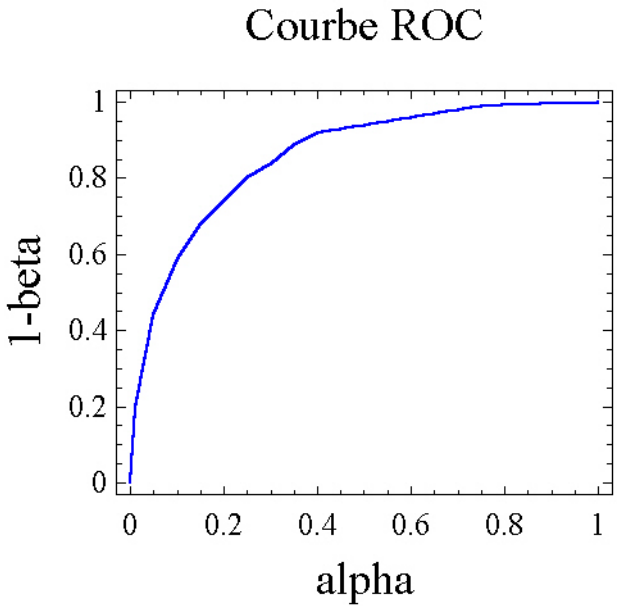
D'une manière similaire à la présentation classique d'un test statistique, la situation peut se décrire à l'aide des deux densités :



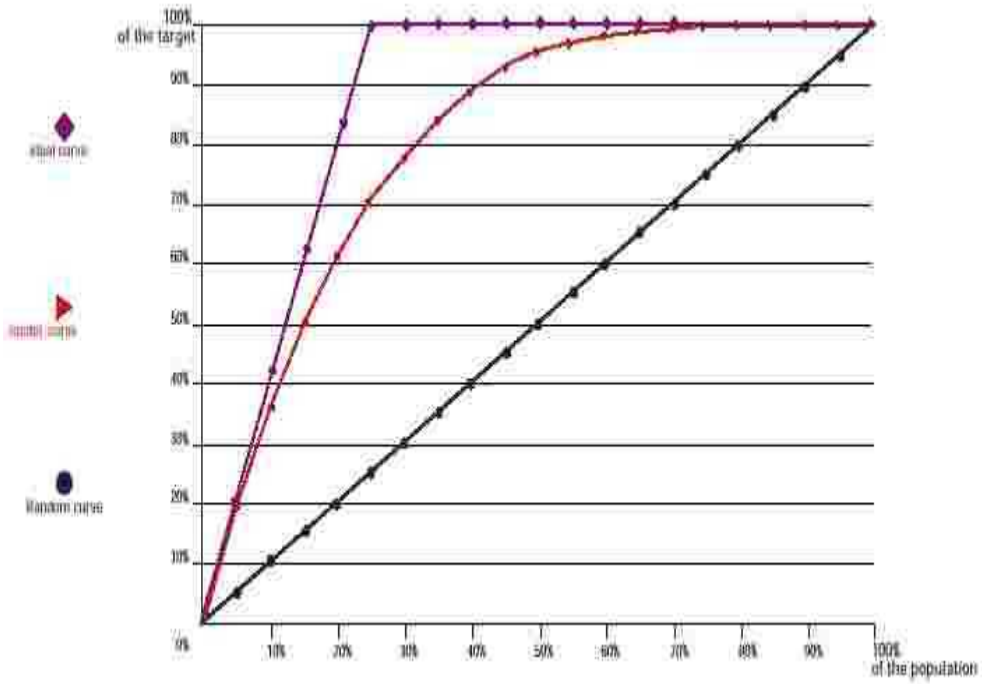
En faisant varier le seuil, on voit qu'en augmentant le pourcentage α de faux mauvais, on augmente aussi le pourcentage 1-β de vrais mauvais.

La courbe suivante (appelée courbe Roc pour " receiver operating curve ") est souvent utilisée pour mesurer le pouvoir séparateur d'un score. Elle donne 1-β(s) en fonction de α(s) lorsque l'on fait varier le seuil s du score. Plus elle est proche de la partie supérieure du carré, meilleure est la séparation. Lorsque les deux densités sont identiques, la courbe ROC se confond avec la diagonale du carré.

La surface entre la courbe et l'axe des abscisses, comprise entre 0 et 1, est également parfois utilisée. On peut montrer qu'elle est théoriquement égale à la probabilité que $P(X_1>X_2)$ si X_1 et X_2 sont deux variables tirées indépendamment, l'une dans la distribution des " bons ", l'autre dans la distribution des " mauvais ".



Les courbes précédentes ne font pas intervenir les proportions réelles de “ bons ” et de “ mauvais ”. Les praticiens utilisent alors la courbe de “ lift ” ou d’efficacité de la



sélection : en abscisse le % de tous les individus bons et mauvais ayant un score inférieur à s, en ordonnée le % de mauvais ayant un score inférieur à s.
La courbe idéale est le segment brisé qui correspond au cas où la distribution des “ mauvais ” est entièrement inférieure à la distribution des “ bons ”.

VI Validité prédictive

Mesurer l’efficacité d’un score, comme d’ailleurs de toute règle de sélection, sur l’échantillon dit “ d’apprentissage ” , c’est à dire celui qui a servi à estimer les coefficients de la fonction de score, conduit à des résultats trop optimistes : en effet les coefficients ayant été optimisés sur cet échantillon, les taux d’erreur sont des estimations biaisées du vrai taux d’erreur, que l’on aura sur de nouvelles données issues de la même population. On peut en effet obtenir de très bons taux de reconnaissance sur l'échantillon d'apprentissage si le nombre de variables explicatives est très élevé : à la limite avec autant de variables que d’observations on pourrait classer sans erreur toute observation, mais ce résultat est purement artificiel.

La validation du score se fait donc à l’aide d’observations supplémentaires, mises de côté, pour lesquelles on connaît Y , et qui servent à simuler le comportement futur du score.

Conclusion

Les méthodes de score, largement utilisées se perfectionnent sans cesse. Elles sont également appliquées dans d'autres domaines : en assurance automobile pour détecter les conducteurs à risque, en prospection publicitaire pour sélectionner des adresses sur un fichier en vue d'un courrier commercial, pour analyser le risque de perte d'un client etc.

Leur usage basé sur une approche statistique permet de mieux quantifier les risques. Bien sur, comme toute méthode statistique, le scoring commet des erreurs et un individu qui a la malchance d'avoir un profil proche de celui de mauvais payeurs sera considéré comme tel ; mais ce type de méthodes commet moins d'erreurs et est plus objectif que les jugements d'expert.

Par ailleurs le score de risque bancaire pour un prêt n'est qu'un élément dans le processus de décision et comme le rappelle la CNIL dans sa Délibération n° 88-083 du 5 juillet 1988 portant adoption d'une recommandation relative à la gestion des crédits ou des prêts consentis à des personnes physiques par les établissements de crédit :
" *conformément à l'article 2 de la loi du 6 janvier 1978, aucune décision accordant ou refusant un crédit ne peut avoir pour seul fondement un traitement automatisé d'informations donnant une définition du profil ou de la personnalité de l'intéressé ".*

Pour en savoir plus :

M Bardos, " Analyse discriminante, application au risque et scoring financier ", Dunod, 2001
Ouvrage de niveau 2ème cycle universitaire, écrit par la responsable de l'Observatoire des Entreprises de la Banque de France. Unique en son genre, en français.

T.Hastie, R.Tibshirani, J.Friedman , " The Elements of Statistical Learning Theory ", Springer-Verlag, 2001
Le livre de référence pour les années à venir, balayant toutes les techniques de modélisation prédictive. Niveau mathématique : 3 ème cycle.

Estimation de courbes de référence pour l'analyse de propriétés biophysiques

Jérôme Saracco[1], Ali Gannoun[1,2] et Christiane Guinot[3]
[1] Laboratoire de Probabilités et Statistique, Université Montpellier II.
Saracco@stat.math.univ-montp2.fr
[2] Statistical Genetics and Bioinformatics Unit, National Genome Center,
Howard University, Washington D. C., U.S.A.
Gannoun@stat.math.univ-montp2.fr
Centre de Recherches et d'Investigations Epidermiques et Sensoriels (CE.R.I.E.S.),
Neuilly-sur-Seine.
Guinot@ceries-lab.com

1 Problématique

De nombreuses expérimentations, en particulier dans le cadre d'études biomédicales, sont conduites pour établir des intervalles de valeurs qui sont prises " normalement " par une variable d'intérêt dans une population cible. Cette variable sera notée Y par la suite. Le terme " normalement " fait référence aux valeurs que l'on est susceptible d'observer avec une probabilité donnée, dans des conditions normales et pour des individus types présumés en bonne santé, ces derniers sont les *sujets de référence*. Ces intervalles sont souvent appelés *intervalles de référence* et les valeurs correspondantes sont appelées *valeurs de référence*. Par exemple, on peut s'intéresser à un intervalle excluant les 5% d'observations les plus grandes et les 5% d'observations les plus petites. Ainsi, la construction d'intervalles de référence repose sur le calcul de quantiles.

D'autre part, il arrive régulièrement que, sur la population cible, l'on dispose simultanément, avec la variable d'intérêt Y, d'une information complémentaire sous la forme d'une covariable X. Très souvent, X représente l'âge du sujet. Pour une valeur donnée x de X, on peut construire un intervalle de référence. Lorsque x varie, on obtient alors des *courbes de référence*. Dans ce cadre, nous sommes amenés à travailler avec les quantiles conditionnels de Y sachant X. Pour les sujets de référence, le tracé de courbes de référence sur le nuage des valeurs prises par le couple (X,Y) donne un résumé graphique très utile et interprétable. Ainsi, un individu i représenté par le point (X_i,Y_i) pourra être comparé à la population de référence. En d'autres termes, une " anormalité " de cet individu sera suspectée si ce point se situe en dessous de la courbe de référence inférieure ou au-dessus de la courbe de référence supérieure.

Plus précisément, pour une valeur x donnée et pour $\alpha \in]0.5, 1[$, l'intervalle de référence contenant $100(2\alpha-1)\%$ des sujets de référence est défini par

$$\mathbf{I}_\alpha(x) = [\, q_{1-\alpha}(x) \,;\, q_\alpha(x) \,],$$

où $q_\alpha(x)$ est le quantile conditionnel d'ordre α de la variable Y sachant que $X=x$. Il est défini de la manière suivante :

$$q_\alpha(x) = F^{-1}(\alpha|x) = \inf\{y \mid F(y|x) \geq \alpha\},$$

$F(.|x)$ désignant la fonction de répartition conditionnelle de Y sachant que $X=x$. Les courbes de référence inférieure et supérieure sont alors les ensembles de points $\{(x,\, q_{1-\alpha}(x))\}$ et $\{(x, q_\alpha(x))\}$ lorsque x varie. En pratique, pour obtenir les courbes de référence à 90%, α est choisi égal à 0,95.

Soit $q_{n,\alpha}(x)$ un estimateur de $q_\alpha(x)$ à partir de l'échantillon $\{(X_i,Y_i),\ i=1,\ldots,n\}$ de n réalisations indépendantes du couple de variables aléatoires (X,Y). L'estimateur correspondant de $\mathbf{I}_\alpha(x)$ est défini par

$$\mathbf{I}_{n,\alpha}(x)=[\ q_{n,1-\alpha}(x)\ ,\ q_{n,\alpha}(x)\ ].$$

Au moins deux types d'approches, l'une paramétrique et l'autre non paramétrique, ont été développés pour l'estimation des quantiles conditionnels et par voie de conséquence des courbes de référence.
L'approche paramétrique repose sur le choix d'une classe paramétrée de distributions. L'estimation des paramètres permet alors de retenir l'une des distributions de cette classe. On a donc forcé la solution à appartenir à une classe de distribution donnée. Ainsi, cette approche paramétrique nécessitant donc des hypothèses restrictives peut être mal adaptée à la réalité des données en particulier biologiques.
L'approche non paramétrique a alors été développée afin de pallier ces problèmes d'hypothèses et de modélisation paramétriques. Les méthodes non paramétriques ne nécessitent en effet pas d'hypothèse sur la nature de la distribution. Elles sont de plus robustes car elles sont déterminées sans détection préalable de points aberrants. En conséquence, une analyse statistique utilisant l'estimation non paramétrique des courbes de référence peut être faite à partir de données d'une fiabilité médiocre.

Dans la suite, nous précisons tout d'abord le cadre et les données de l'étude. Nous détaillons ensuite la méthode non paramétrique d'estimation par noyau des quantiles conditionnels et nous appliquons cette méthode pour construire les courbes de référence à 90% d'un paramètre biophysique de la peau. Enfin, nous terminons en mentionnant quelques extensions de ce travail, en particulier, au cadre multidimensionnel et à d'autres estimateurs non paramétriques des quantiles conditionnels.

2 Les données

En vue de cibler des produits cosmétiques sur le marché japonais, Chanel a demandé au CE.R.I.E.S. (centre de recherche sur la peau humaine financé par Chanel et situé à Neuilly-sur-Seine) de faire une étude sur les propriétés biophysiques de la peau de femmes japonaises. La Statistique, au moyen de l'estimation de courbes de référence en fonction de l'âge pour ces propriétés biophysiques, va ainsi servir d'aide à la décision pour adapter au mieux les produits à ce nouveau marché asiatique.
L'objectif de l'étude (partiellement présentée ici) réalisée par le CE.R.I.E.S. était d'établir donc ces courbes de référence à 90% pour les propriétés biophysiques (mesurées sur deux zones du visage et une zone de l'avant-bras).

Les données utilisées ont été recueillies par le CE.R.I.E.S. entre le 15 décembre 1998 et le 15 avril 1999 à Sendai (Japon) sur n=120 femmes japonaises présentant une peau apparemment saine (c'est-à-dire sans aucun signe de dermatose en cours ou de maladie générale avec manifestations cutanées avérées). Chaque volontaire a été examinée en atmosphère contrôlée (température de 23 ± 1°C et humidité relative de 50 ± 5%). Cette étude comportait des questionnaires sur les habitudes de vie, un interrogatoire et un examen médical cutané, ainsi qu'une évaluation des propriétés biophysiques cutanées. L'évaluation des paramètres biophysiques a été effectuée sur deux zones du visage (front et joue) et sur la face antérieure de l'avant-bras gauche. Les paramètres biophysiques (variables d'intérêt) mesurés ou calculés

sont les suivants : le taux de sécrétion de sébum (mesuré uniquement sur le front et la joue), la température cutanée, la perte insensible en eau, le pH cutané, l'hydratation de la peau par capacitance et conductance, la couleur de la peau (exprimée à l'aide de trois grandeurs), l'angle typologique individuel, la saturation et l'angle de teinte. La covariable est l'âge des volontaires.

Dans la suite de ce document, nous nous limiterons uniquement à l'étude de la variable mesurant le sébum instantané de la joue (en µg/cm^2), notée SJOUE, en fonction de l'âge.

3 Un peu de modélisation mathématique

Rappelons tout d'abord qu'un estimateur de la fonction de répartition conditionnelle de Y sachant $X=x$, est défini, pour $y \in \boldsymbol{R}$, par :

$$F_n(y|x)= \sum_{i|X_i=x} \left(\frac{1}{n_x}\right) I(Y_i \le y) \quad \text{si } n_x>0, \quad \text{et} \quad F_n(y|x)=0 \text{ sinon,}$$

où n_x désigne le nombre d'individus de l'échantillon tels que $X_i=x$, et $I(.)$ est la fonction indicatrice. Il apparaît clairement que, pour faire une estimation correcte de cette fonction de répartition, il faut disposer d'un certain nombre d'observations Y_i telles que $X_i=x$, ce qui est rarement le cas en pratique. De plus, cet estimateur n'est pas " lisse " (continue) en x. Il est donc préférable d'introduire un estimateur qui permet de contourner ces problèmes.

Définissons alors maintenant un estimateur non paramétrique (" lisse " en x) de la fonction de répartition conditionnelle de Y sachant $X=x$, pour $y \in \boldsymbol{R}$:

$$\widetilde{F}_n(y|x)= \sum_{i=1}^{n} \left(\frac{K\{(x-X_i)/h_n\}}{\sum_{j=1}^{n} K\{(x-X_j)/h_n\}} \right) I(Y_i \le y) .$$

La fonction K, appelée noyau[2], permet de faire intervenir dans le calcul de l'estimateur tous les points de l'échantillon affectés de poids d'autant plus grand que X_i est voisin de x. On prend généralement pour K la densité de la loi normale centrée réduite. Le paramètre h_n, appelée fenêtre, permet de contrôler le lissage appliqué aux données : plus h_n est grand, plus l'estimateur va prendre en compte un nombre important d'observations et donc plus le lissage sera important ; à l'inverse plus h_n est petit, et moins l'estimateur sera " lisse ". Son choix est crucial en pratique car il faut éviter de faire du sur-lissage ou du sous-lissage. Un choix " optimal " pour h_n peut être obtenu de manière automatique à partir des données au moyen d'une méthode de validation croisée[3].

De l'estimateur de la fonction de répartition conditionnelle $\widetilde{F}_n(y|x)$, il est alors naturel d'estimer le quantile conditionnel $q_\alpha(x)$ par $\widetilde{q}_{n,\alpha}(x)$ défini de la manière suivante :

$$\widetilde{q}_{n,\alpha}(x) = \widetilde{F}_n^{-1}(y|x) = \inf\{y \mid \widetilde{F}_n(y|x) \ge \alpha\},$$

l'inversion étant faite de manière numérique.

---

[2] fonction que l'on suppose généralement positive, symétrique et maximale en zéro (très souvent, il s'agit d'une densité de probabilité)

[3] Pour plus de détails, nous renvoyons le lecteur aux deux articles (1) et (2) mentionnés en bibliographie.

4 Application à la variable SJOUE

Avant de fournir les résultats obtenus pour cette variable d'intérêt, on précise que les courbes de référence obtenues sont considérées comme acceptables si elles satisfont les trois conditions suivantes :
(a) Elles n'incluent pas de valeurs impossibles pour Y (*i.e.* par exemple, des valeurs nulles ou négatives alors que la variable Y ne peut prendre en réalité que des valeurs strictement positives).
(b) Elles contiennent le pourcentage désiré d'individus à savoir ici 90%.
(c) Les valeurs individuelles qui se trouvent en dehors des limites des courbes de référence sont réparties de façon uniforme en fonction de la covariable AGE et aucun regroupement de valeurs individuelles n'apparaît.

Trois méthodes non paramétriques d'estimation des quantiles conditionnels et donc des courbes de référence ont été mise en oeuvre dans cette étude : la méthode d'estimation par noyau décrite précédemment, ainsi qu'une méthode d'estimation par noyau dite de la constante locale et une méthode d'estimation par noyau produit[4]. La Figure 1 nous donne le nuage des points croisant les variables AGE et SJOUE sur lequel les courbes de référence à 90% obtenues avec les différentes méthodes non paramétriques ont été superposées.

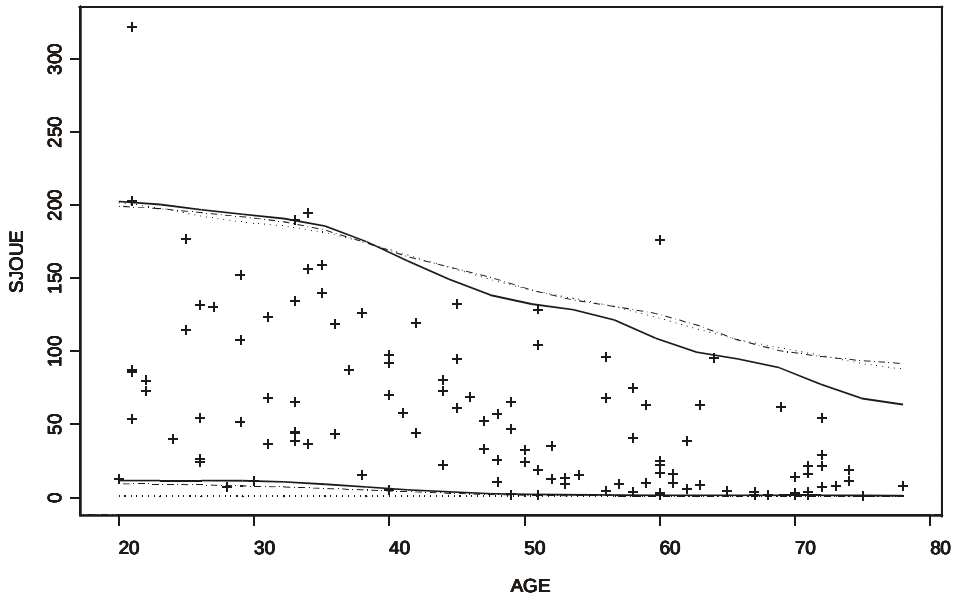


Figure 1 : *Courbes de référence à 90% obtenues avec des méthodes non paramétriques pour la variable SJOUE (trait continu : méthode d'estimation par noyau, pointillés : méthode de la constante locale, tirets : méthode d'estimation par noyau produit).*

On peut constater que ces courbes de référence sont physiologiquement acceptables. En particulier les courbes de référence supérieures obtenues avec les trois estimateurs non paramétriques correspondent bien à ce que l'on s'attend à observer d'un point de vue biologique (décroissance du taux instantané de sébum avec l'âge), seul l'aspect légèrement

[4] Pour des références sur ces différentes méthodes, le lecteur pourra consulter les articles (1) et (2) de la bibliographie.

“ ondulé ” n'est pas totalement conforme. Il serait cependant “ techniquement ” possible de “ lisser ” un peu plus ces courbes en prenant des fenêtres légèrement plus larges que celles sélectionnées automatiquement par la méthode de validation croisée.

5 Compléments et extensions

Bien qu'extrêmement utiles, les estimateurs non paramétriques mentionnés ici ne portent que sur une seule variable d'intérêt et ne prennent en compte qu'une seule covariable (l'âge dans notre application). Il serait intéressant de les adapter à des situations plus générales.

(*i*) D'une part, il est parfois nécessaire d'utiliser plusieurs covariables quantitatives (l'âge, le poids et les conditions expérimentales par exemple) pour augmenter la précision des courbes de référence. Le fait que la covariable devienne multidimensionnelle, n'est pas un obstacle au développement de la théorie. Cependant les méthodes non paramétriques souffrent d'un point de vue pratique du “ fléau de la dimension ”[5]. Se pose aussi le problème de la représentation graphique des données et des estimateurs. Malgré l'évolution des logiciels, il est impossible de faire des représentations graphiques sérieuses quand la dimension de l'espace qui contient les variables est supérieure à 3. Ceci prive l'analyste de faire des constatations vraisemblables et d'avancer des conclusions plausibles à la seule lecture des graphiques.

Nous avons alors proposé une nouvelle méthodologie fondée sur une étape de réduction de la dimension de la covariable, suivie d'une étape d'estimation non paramétrique de quantiles conditionnels. Cette approche semiparamétrique combine la méthode SIR (Sliced Inverse Regression) et l'estimation à noyau de quantiles conditionnels. L'étape de réduction de la dimension permet d'obtenir un ou plusieurs indices $\beta' X$ résumant la partie explicative de la covariable vectorielle X. Notons que cette technique de réduction de la dimension au moyen d'indices entraîne non seulement l'amélioration de l'estimation du quantile conditionnel (disparition ou forte diminution du fléau de la dimension), mais donne aussi aux analystes la possibilité d'interpréter et de quantifier le rôle joué par chaque covariable dans l'étude. Nous avons illustré cette méthode en établissant des courbes de référence pour des propriétés biophysiques de la peau de femmes françaises “ saines ” en fonction de l'âge, des conditions expérimentales et d'autres propriétés biophysiques.

(*ii*) D'autre part, le cadre où la variable d'intérêt est multidimensionnelle intéresse aussi les praticiens, en particulier ceux travaillant sur la peau. On parle alors non plus de courbes de référence mais de “ régions ” de référence multivariées. Par exemple, la variable d'intérêt multidimensionnelle aurait pour composantes toutes les paramètres biophysiques évalués sur une zone particulière (la joue, le front ou l'avant-bras).
En effet, bien que la valeur des paramètres de certains individus soient dans les limites de référence des paramètres pris isolément, lorsque les valeurs de ces mêmes individus sont examinées grâce à des méthodes multivariées, ils peuvent apparaître en dehors de la “ région ” de référence. Cette apparente contradiction est généralement due aux corrélations entre les différents paramètres biophysiques qui ne sont pas prises en compte avec les courbes de référence (univariées). Ces corrélations pourraient ainsi être introduites dans la construction de “ régions ” de référence multivariées, ce qui en fait leur intérêt majeur. Par exemple,

[5] Cela décrit le fait qu'il devient de plus en plus difficile de faire une estimation de bonne qualité vu la dispersion de plus en plus grande des données dans un espace de grande dimension.

lorsque pour un sujet la corrélation entre les paramètres n'est pas conforme à celle attendue, compte tenu des informations apportées par la totalité de l'échantillon, une anomalie multivariée apparaît, anomalie qui ne pourrait pas être détectée par l'examen isolé de chacun des paramètres. La révélation de cette anomalie permet l'identification de sujets particuliers, et de procéder à un examen minutieux des données de ces individus.
L'estimation des quantiles conditionnels multivariés permet d'élargir l'approche non paramétrique à ce contexte multidimensionnel.

(*iii*) Plus généralement, lorsque la variable d'intérêt et la covariable sont toutes les deux multidimensionnelles, une extension naturelle est la combinaison de ces deux directions. Ce thème est actuellement en cours d'étude.

Références bibliographiques

Pour plus de détails théoriques et des résultats plus détaillés concernant cette étude et une étude similaire faite sur un échantillon de femmes françaises pour lesquelles les courbes de référence à 90% ne sont établies qu'en fonction de l'âge, le lecteur pourra se référer aux deux premiers articles cités ci-après. En ce qui concerne la construction de courbes de référence utilisant une covariable multidimensionnelle, nous renvoyons le lecteur au troisième article.

(1) Gannoun, A., Girard, S., Guinot, C. & Saracco, J. (2002). " Trois méthodes non paramétriques pour l'estimation de courbes de référence - Application à l'analyse de propriétés biophysiques de la peau ". *Revue de Statistique Appliquée*, **L**(1), 65-89.

(2) Gannoun, A., Girard, S., Guinot, C. & Saracco, J. (2002). " Reference curves based on nonparametric quantile regression ". *Statistics in Medicine*, **21**, 3119-3135.

(3) Gannoun, A., Girard, S., Guinot, C. & Saracco, J. (2001). " Dimension reduction in reference curves estimation ". *Rapport technique ENSAM-INRA-UM2,* n° 01-06.

Les traitements statistiques de données textuelles.

Ludovic Lebart
CNRS-ENST
lebart@enst.fr

Le matériau statistique " texte " est omniprésent, presque banal, depuis le développement d'internet et de la toile (*web*). L'étude quantitative et statistique de ces textes semble avoir fait irruption récemment, et pourtant les études statistiques de textes datent de plusieurs décennies, avec notamment en France les travaux de P. *Guiraud (Problèmes et méthodes de la statistique linguistique*, PUF, 1960), C. Muller (*Principes et méthodes de statistique lexicale*, Hachette, 1977) puis de J.P. Benzécri (Pratique de l'Analyse des Données, Tome 3 : *Linguistique et lexicologie*, Dunod, 1981).

Après la " stylométrie ", consacrée à l'étude de la forme des textes, en vue d'identifier un auteur ou de dater une œuvre, sont apparues les techniques de documentation automatique (*Information retrieval* en Anglais), visant à rechercher dans une base de documents (articles scientifiques, résumés, brevets, …) le ou les éléments pertinents à partir d'une requête exprimée sous forme de textes libres. Le champ disciplinaire " Traitement du Langage Naturel " est alors apparu, et s'est développé, au départ, comme un des domaines d'application privilégié de l'intelligence artificielle. La complexité du matériau, le besoin d'assimiler d'immenses corpus de textes, la pertinence du concept d'apprentissage ont naturellement ouvert ce champ aux méthodes statistiques. La statistique multidimensionnelle, les chaînes de Markov cachées, les méthodes d'analyse discriminantes interviennent ainsi pour construire les outils de base que sont les moteurs de recherche sur le *web*, les analyseurs morphosyntactiques, les correcteurs orthographiques, ainsi que dans des champs d'application pratiques comme le traitement des réponses aux questions ouvertes dans les enquêtes socio-économiques.

Les questions ouvertes

Il est utile, dans un certain nombre de situations d'enquête, de laisser ouvertes certaines questions, dont les réponses se présenteront donc sous forme de textes de longueurs variables.

Dans au moins trois situations courantes, l'utilisation d'un questionnement ouvert s'impose :

Pour diminuer ou optimiser la durée de l'entrevue d'enquête

Bien que les réponses libres et les réponses guidées fournissent des informations de natures différentes, les premières sont plus économiques que les secondes en temps d'interview et génèrent moins de fatigue. Une simple question ouverte (par exemple : "Quelles furent vos principales activités dimanche dernier ?") peut remplacer de longues listes d'items.

Comme complément à des questions fermées

Il s'agit le plus souvent de la question: *"Pourquoi ?"*. Les explications concernant une réponse déjà donnée doivent nécessairement être spontanée. Une batterie d'items risquerait de proposer de nouveaux arguments qui pourraient nuire à l'authenticité de l'explication. L'utilité de la question *pourquoi ?* a été soulignée par de nombreux auteurs, et ce sont en fait les difficultés et le coût de l'exploitation qui en limitent l'usage. Elle seule permet en effet de savoir si les différentes catégories de personnes interrogées ont compris la question fermée de la même façon.

Pour recueillir une information qui doit, par nature, être spontanée

Les questionnaires des enquêtes de marketing abondent en questions de ce type. Citons par

exemple : "Qu'avez-vous retenu de cette campagne publicitaire ?", "Que pensez-vous de cette voiture ?". Notons cependant que les questions ouvertes sont considérées comme peu adaptées aux problèmes de mémorisation de comportement. "Quels magazines avez-vous lus la semaine dernière ?", "Quelles sont les dernières émissions de télévision que vous avez aimées ?". Pour ces questions qui font l'objet d'enquêtes périodiques, il a été prouvé maintes fois que les questions fermées donnent des taux d'oubli plus faibles. En revanche, quand la qualité de la mémorisation est en jeu, la forme ouverte reste indispensable.

Voici quatre exemples de réponses à la question " Quelle est pour vous la chose la plus importante dans la vie ? " (Question posée à des échantillons d'environ mille personnes dans sept pays en1991).

1) La santé, ne pas manquer d'argent, avoir une bonne ambiance familiale, je voudrais pouvoir aider les enfants abandonnés, leur redonner le goût à la vie, pouvoir aider les personnes âgées handicapées, secourir les gens autour de soi.

2) C'est de faire ce qu'on veut. Lire, voyager si je pouvais. Les loisirs si on pouvait.

3) La santé puisqu'il faut toujours travailler quand on est commerçant. Une bonne entente en famille. Avoir assez d'argent pour vivre.

4) La famille, ma famille, mon foyer, vivre avec la société : mon entourage les voisins, pour faire quelque chose qu'il y ait moins de malheureux, donner du travail aux jeunes surtout.

Ces exemples illustrent à la fois la complexité et la richesse des réponses.

Les unités statistiques

Les programmes travaillent à partir du texte brut, en extrayant automatiquement des unités statistiques, la plupart du temps des *formes graphiques* (séquences de caractères non-séparateurs). On utilise le vocable *forme graphique* parce que le mot " mot " lui-même est ambigu. Il désigne en effet selon les contextes *l'occurrence* d'un mot (quand on dit qu'un texte a huit cent mots, on parle bien sûr d'occurrences, et non de mots différents), le type (qui correspond à la forme graphique) et le *lemme* (*avoir* est le lemme de *avait*, et, dans certains cas seulement, de *avions*). La première réponse de l'exemple ci-dessus contient 38 occurrences, mais la forme graphique " les " apparaît trois fois, " pouvoir " apparaît deux fois. Le lemme de " bonne " est bon (le masculin singulier, selon une convention française), celui de " voudrais " est " vouloir ".

Dans le cas de l'exemple précédent, pour 1009 réponses, on obtient 14337 occurrences de 1394 formes distinctes (ou types). Il est bien connu que la distribution de fréquence des mots est très dissymétrique (loi dite de Zipf, apparentée à la distribution de Paréto). Ainsi, en ne retenant que les formes apparaissant au moins 20 fois, il reste un texte de 10 994 formes, avec seulement 97 formes distinctes (ainsi 7 % des mots distincts correspondent à 77 % du texte global). En particulier, près de la moitié des formes grahiques distinctes n'apparaissent qu'une fois (ce sont les " hapax ").

Le post-codage

Le prétraitement empirique appelé "post-codage" permet de fermer *a posteriori* les questions ouvertes. Cette technique courante consiste à construire une batterie d'items à partir d'un sous-échantillon de réponses, puis à codifier l'ensemble des réponses de façon à remplacer la question ouverte par une ou plusieurs questions fermées. Pour l'exemple ci-dessus, la seconde réponse, la plus simple, donnerait les items " lecture ", voyage ", " loisirs ", sous réserve que ces items apparaissent avec une certaine fréquence dans l'échantillon de réponses. En revanche la première réponse est plus délicate à post-coder.

Les outils statistique de base

Les outils de base sont la sélection de formes caractéristiques, la sélection de réponses modales, l'analyse des correspondances et la classification des tableaux lexicaux**.**

Formes ou segments caractéristiques (ou spécificités)

Les formes caractéristiques sont les formes "anormalement" fréquentes dans les réponses d'un groupe d'individus (technique proposée par P. Lafon en 1980). Un test élémentaire fondé sur la loi hypergéométrique permet de sélectionner les mots (formes graphiques ou lemmes) dont la fréquence dans un groupe est notablement supérieure (ou inférieure pour les mots *anti-caractéristiques)* à la fréquence moyenne dans le corpus. Il s'agit de test classique de comparaisons de fréquences, mais la répétition de ce test conduit à prendre des seuils de signification très sévères (phénomène de *comparaisons multiples* bien connu des statisticiens).

Dans l'exemple évoqué plus haut, la fréquence moyenne du mot travail dans le corpus était de 3.4 %; pour le groupe des femmes de plus de 55 ans, la fréquence n'est que de 1.2 %. Cette différence est en fait hautement significative (on peut exprimer le test de comparaison de fréquences en termes d'écart-types : dans l'hypothèse d'homogénéité des fréquences, la valeur1.2% est à 4.5 écart-types de la valeur moyenne 3.4). Comme il s'agit d'une fréquence anormalement faible, on parlera de mots anti-caractéristiques. [L'individu statistique est ici l'occurrence de mots. Les femmes de plus de 55 ans ont émis 1349 mots dans leurs réponses. La variance de la fréquence d'un mot dont la frequence "théorique" est de 0.034 est donnée par la formule classique $0.034(1 - 0.034) /1349$. On voit dans ces conditions qe la fréquence observée de 0.012 est à 4.5 écart-types de 0.034].

Les sélections des réponses modales

Pour un groupe d'individus donné, et donc pour le regroupement de réponses correspondant, les réponses modales (ou encore phrases caractéristiques, ou documents-type, la terminologie variant selon les domaines d'application) sont des réponses originales du corpus de base, ayant la propriété de caractériser au mieux le groupe. On peut, pour chaque regroupement, calculer la distance du profil lexical d'un individu au profil lexical moyen du groupement. On peut ensuite classer les distances par ordre croissant, et donc sélectionner les réponses les plus représentatives au sens du profil lexical, qui correspondront aux plus petites distances. On obtient ainsi une sorte de résumé des réponses de chaque regroupement, formé de réponses originales (L. Lebart et A. Salem, *Statistique Textuelle*, Dunod, 1994). Toujours dans le cas de notre exemple, "Etre heureux, avoir un bon travail, réussite professionnelle et familiale" est ainsi une réponse caractéristique des jeunes hommes; "la santé, la famille" est une réponse caractérisant les plus âgés. On utilise en pratique plusieurs réponses caractéristiques par groupe.

Analyse des correspondances et classification

Le volume des données demande que l'on fasse appel à de puissants outils de description. Les méthodes d'analyses des correspondances et de classification peuvent décrire les tables de contingence croisant les réponses et les formes graphiques, ou des groupes de réponses (par exemple regroupement selon le niveau d'instruction des répondants) et les formes graphiques. Elles permettent de visualiser sous forme de séries de cartes planes (ou de dendrogrammes dans le cas des méthodes de classification, ou de *cartes auto-associatives* de Kohonen, méthode "neuronale" de visualisation) les associations entre mots (formes) et groupes ou modalités. Ainsi, une visualisation des proximités entre mots et catégories socioprofessionnelles pourra aider la lecture des réponses de chacune de ces catégories.

Conclusions et ouvertures

Pour des réponses simples et stéréotypées, nous l'avons vu, les procédures de post-codage peuvent fonctionner. Mentionnons cependant parmi les défauts de ce type de traitement :

La médiation du chiffreur: les décisions à prendre sont parfois difficiles.

La qualité de l'expression, le registre du vocabulaire, la tonalité générale de l'entretien sont des éléments d'analyse perdus lors d'un post-codage (doit-on coder différemment " je ne sais pas" et "je préfère ne rien dire" ?.

Les réponses composites, complexes, d'une grande diversité, sont très difficile à post-coder, et c'est souvent dans ce cas que la valeur heuristique des réponses libres est la plus grande.

Les réponses peu fréquentes, originales, peu claires en première lecture sont considérées comme du "bruit", et affectées à des items résiduels ("autres") qui sont donc très hétérogènes et sont difficiles à manipuler.

Sans qu'il soit nécessaire de procéder à un post codage, on peut, actuellement, à partir d'un ensemble de textes, et d'un seuil de fréquence pour les formes graphiques, obtenir une visualisation des proximités entre textes (vis-à-vis de leurs profils lexicaux) et entre formes graphiques (vis-à-vis de leur répartition dans les textes). L'enrichissement des unités statistiques par les *segments répétés* (cf. A. Salem, *Pratique des segments répétés,* Klincksieck, 1987), leurs regroupements par catégorisation morphologique, l'utilisation des formes caractéristiques ou spécificités, l'adjonction des réponses modales ou des phrases ou unités de contexte caractéristiques ont perfectionné ces approches, et mis à la disposition de beaucoup d'utilisateurs des méthodes et des logiciels utiles. Dans certains domaines d'application précis (comme le traitement automatique des réponses aux questions ouvertes, qui nous intéresse ici), l'efficacité de la méthode*, comme complément des approches traditionnelles*, est reconnue.

Parallèlement aux travaux relevant de *l'Industrie de la Langue*, que nous avons évoqués plus haut, et qui relèvent d'une *ingéniérie statistique* complexe, il existe donc des applications textuelles de la statistique qui restent à portée de main. Elles nécessitent certes des logiciels spécifiques, mais la nature familière et vivante du matériau de base compense en quelque sorte la relative complexité des traitements et les difficultés d'interprétation.

Proche des bases de données, de l'intelligence artificielle et des réseaux de neurones, de la théorie de l'apprentissage, des techniques récentes d'extraction et de gestion des connaissances, le domaine textuel illustre bien la polyvalence et la puissance de la méthodologie statistique. Même quand les méthodes prennent parfois les noms plus exotiques de *fouille de texte* ou de *text mining*, le statisticien est toujours sollicité quand il s'agit de connaître la portée réelle des faits observés et des traits structuraux obtenus, de savoir ce que l'on a le droit de dire ou le devoir de ne pas dire, c'est-à-dire finalement de donner un statut scientifique aux résultats.

Sensibilisation à la Statistique

Yves Escoufier
Laboratoire de Probabilités et Statistiques
Université Montpellier 2
yves.escoufier@univ-montp2.fr

Le texte qui suit donne la trame d'une séance de sensibilisation à la Statistique faite à l'invitation de l'IUFM et de l'IREM de Dijon pour des professeurs et élèves professeurs de mathématiques des lycées et collèges de l'académie de Bourgogne. L'objectif annoncé aux auditeurs était de parcourir les grandes étapes d'une démarche d'analyse statistique en tentant pour chacune d'entre elles d'inviter à des prolongements personnels.

I) Un thème d'étude et des données

La proposition faite aux auditeurs est de prendre pour thème d'étude la morphologie de la main gauche. Dans ce but, chacun des participants est invité à placer sa main gauche sur une feuille de papier, bien à plat et doigts écartés le plus possible, à tracer le contour de la main avec un crayon et à marquer dans un coin de la feuille H ou F selon qu'il est de sexe masculin ou féminin.
Plusieurs commentaires peuvent être faits à ce moment pour dépasser le thème d'étude lui même . On peut espérer que l'intérêt que les auditeurs portent au thème de l'étude alimentera l'intérêt qu'ils manifesteront pour les éléments de Statistique qui vont être présentés. En ce sens , le choix du thème n'est pas neutre : plus il concerne les auditeurs, plus grandes sont les chances de retenir leur attention pour la Statistique. L'aspect en partie ludique du tracé du contour de la main est aussi un élément susceptible de faciliter l'intérêt. Il faut cependant être prudent : quelle serait ici la réaction d'un auditeur dont la main gauche serait déformée ? Il peut y avoir des réactions négatives au thème proposé ; elles ne doivent pas être négligées.
On peut imaginer que des statisticiens expérimentés s'engagent directement dans l'étude des courbes particulières que sont les contours des mains recueillis. Ce n'est bien sûr pas le cas dans cette séance de sensibilisation. On va donc devoir choisir quelques caractéristiques qui deviendront les objets des études statistiques ultérieures. Des kinésithérapeutes, des ergonomes, des chirurgiens spécialisés dans les soins de la main auraient à coup sûr des suggestions à faire pour le choix de ces caractéristiques : ils ont des connaissances sur le sujet ; ils connaissent les résultats d'études antérieures qui pourraient servir de guide. En bref, l'étude qui sera faite, comme toute étude, ne portera que sur des aspects restreints du phénomène étudié ; la restriction dépend pour une part de nos connaissances antérieures sur le sujet , pour une autre part des outils de mesure et d'enregistrement des données dont nous disposons et de nos capacités à les maîtriser.
Pour l'étude présente, nous retiendrons une mesure de la largeur de la paume et une mesure de la longueur du majeur. A ces deux caractéristiques numériques, s'ajoute la caractéristique qualitative " sexe " dont les deux valeurs sont " homme " et " femme ". On peut ici élargir l'exposé aux différents types de caractéristiques (numériques continues, numériques discrètes, ordinales, qualitatives, dichotomiques), passer du temps sur le concept de variable statistique, sur ses valeurs ou modalités, sur ses fréquences. Une ouverture peut être faite sur les variables statistiques multidimensionnelles.
Les lancers d'une pièce de monnaie ou d'un dé sont souvent pris comme thèmes d'études. Ils permettent effectivement sans grand investissement de disposer de données qui peuvent

interpeller des élèves et donc faciliter leur attention en faveur des éléments de Statistique qu'ils permettront d'introduire. Ils présentent un autre avantage. Le nombre des expériences qui peuvent être faites avec une pièce de monnaie ou un dé reste limité. Il devient alors assez naturel de rechercher un procédé technologique qui réagisse comme un dé ou une pièce de monnaie. Le simulateur de nombres uniformes sur (0,1) s'introduit ici naturellement et il est facile en commençant par la simulation de dés pipés ou de pièces truquées de montrer qu'il va permettre de simuler des expériences complexes.

II) La descriptions des données

Afin de permettre aux auditeurs de reproduire eux mêmes les démarches qui vont être faites, le choix est fait d'utiliser " l'utilitaire d' analyse " que l'on trouve dans la rubrique " outil " du logiciel Excel ainsi que la fonction " graphique " de la rubrique " insertion " . Les données recueillies sont données dans le tableau 1. Elles permettent de mettre en œuvre un certain nombres de méthodes de description de données offertes par le logiciel, d'évoquer les problèmes implicites éventuels de leur mise en œuvre, de commenter les résultats.

- Le tracé d'un histogramme pose le problème du choix des classes ;
- Il est intéressant de vérifier comment le logiciel calcule effectivement la médiane, les quartiles et plus généralement les quantiles ;
- Obtenir des graphiques agréables à lire suppose de réaliser des translations des données ;
- Le logiciel propose un écart type résultant d'une division par n et un autre que l'on ne peut que renvoyer à plus tard divisé par n-1 ;
- Les analyses faites sur l'ensemble des données peuvent facilement être reprises pour les hommes seuls ou les femmes seules : voilà une bonne introduction pour l'adjectif " conditionnel " en Statistique.

A titre d'illustration, les tableaux 2 et 3 donnent les histogrammes de la variable " paume " pour les femmes et pour les hommes. Les classes ont été choisies identiques pour les deux groupes. Les graphiques mettent bien en évidence que pour les données recueillies, les paumes les plus larges sont masculines et les paumes les moins larges féminines.

III) Statistique et Probabilités

S'il ne s'agissait que de décrire les cinquante mains observées, nous pourrions nous arrêter ici. Mais le statisticien veut le plus souvent étendre à des données potentielles non observées les résultats de ses observations limitées. Pour cela, il a besoin d'un cadre qui le guide et donne un sens aux données partielles qu'il recueille. La théorie des probabilités va le lui fournir.

La question est parfois posée de savoir pourquoi ce sont les enseignants de mathématiques qui doivent enseigner les statistiques dans les lycées. La réponse vient de la nécessité d'utiliser les concepts et les résultats des probabilités, branche des mathématiques, pour donner un sens à la démarche statistique.

En effet, toute extension des résultats obtenus sur un fini observé à un ensemble plus grand qui l'englobe, suppose que l'observé puisse être considéré comme une réalisation d'un échantillon représentatif de l'ensemble plus grand. On ne peut échapper à la théorie des probabilités pour donner un sens à ces mots : échantillon, représentatif, réalisation.

Un enseignement classique devrait traverser les étapes de l'espace probabilisable et des opérations sur les événements ; il continuerait par une présentation axiomatique des probabilités avec éclairage sur les probabilités conditionnelles et les événements indépendants ; viendraient alors l'espace probabilisé, la notion de variable aléatoire et le

théorème de transfert de la probabilité. A ce stade l'étude des modèles classiques de variables aléatoires peut être faite ce qui introduit dans le vocabulaire les mots de " lois et distributions de probabilité ".Le concept de variables aléatoires indépendantes conduit alors à la notion d'échantillon de variables aléatoires indépendantes de même loi. Reste ensuite à parler des réalisations d'un échantillon puis des fonctions définies sur un échantillon et des valeurs prises par ces fonctions sur les réalisations de l'échantillon.
Bien sûr, on ne peut pas faire tout cela de manière rigoureuse dans le secondaire. Mais celui ou celle qui l'enseigne doit le posséder d'autant mieux qu'il doit en donner une version simplifiée mais juste. La simulation peut – elle aider dans cet enseignement ? Certainement. La simulation d'une certaine loi peut être assimilée à la variable aléatoire qui obéit à cette loi ; le résultat de chacun des appels du simulateur est alors considéré comme une réalisation de cette variable. On peut donc visualiser les fluctuations de ces réalisations mais aussi calculer des fonctions de ces réalisations telles que la moyenne ou la variance. Cette approche expérimentale se substitue à l'étude analytique des variations de la fonction de densité.
Dans le même esprit, n appels successifs du simulateur peuvent être assimilés à un échantillon de taille n de la variable qu'il simule. Une réalisation de ces n appels fournit une réalisation de l'échantillon de taille n. Les fluctuations de toutes les fonctions de ces réalisations peuvent être étudiées en répétant les n appels.
Nous venons de dire que la Statistique avait besoin des probabilités : elles lui donnent son vocabulaire, ses concepts et les propriétés des objets qu'elle manipule . En retour, par son contact avec les différents domaines dans lesquels elle intervient, la Statistique est sans cesse confrontée à des problèmes nouveaux et à des données de types nouveaux. Elle demande aux probabilités de construire les outils dont elle a besoin et d'en préciser les propriétés. Par là , la Statistique alimente le développement des probabilités. Pour donner deux exemples actuels, évoquons les données spatiales et celles liées au génome. Le statisticien les a rencontrées avant d'avoir les outils adaptés à leur étude. Peu à peu avec les probabilistes, il dégage les concepts nécessaires et met en évidence leurs propriétés.

IV) Estimation et tests

Revenons à l'étude proposée en I. Les mesures ont été faites sur des femmes et des hommes adultes, étudiants ou enseignants de mathématiques. Peut–on d'une manière ou d'une autre étendre les résultats constatés sur les quinze femmes et les trente cinq hommes observés à l'ensemble des femmes et hommes adultes , étudiants ou enseignants de mathématiques ? C'est là tout le champ de l'inférence statistique.
Rien ne pourra être fait sans pouvoir considérer que , par exemple, les quinze femmes étudiées constituent un échantillon représentatif de la population des femmes adultes, étudiantes ou enseignantes en mathématiques ce qui veut dire que les quinze mesures obtenues sont bien une réalisation d'un échantillon de taille quinze de la variable paume pour cette population. C'est donc la procédure de constitution de l'échantillon qui doit être interrogée. Il y a des livres entiers sur ce sujet : échantillons aléatoires ; échantillons stratifiés ; méthodes des quota. Nous n'irons pas plus loin ici sur ce thème.
Cette question traitée, l'inférence statistique ouvre sur deux domaines : l'estimation et les tests.
L'estimation consiste à induire des observations faites sur l'échantillon des informations valables pour la population. Dans sa forme dite paramétrique, elle consiste à postuler le modèle qui régit la variable aléatoire objet de l'étude (on postulera par exemple qu'elle suit une loi normale) et à obtenir de l'échantillon des estimations des paramètres qui déterminent cette loi (la moyenne et la variance pour la loi normale). Ce faisant, on a ajouté aux données disponibles, l'hypothèse d'un modèle pour la variable aléatoire. Il faut comprendre l'intérêt

que représente une telle association d'un modèle théorique à une variable. Si l'association est possible (condition qui sera mise à l'épreuve d'un test), tout ce qui est connu pour le modèle devient applicable à la variable.
Plusieurs procédures sont disponibles pour estimer les paramètres. Le sujet est complexe. Par exemple la méthode dite du maximum de vraisemblance conduit au diviseur n pour l'écart type mais lorsqu'on étudie les propriétés de cet estimateur on découvre qu'il est biaisé, c'est à dire que la moyenne de cet estimateur sur l'ensemble des échantillons envisageables n'est pas la vraie valeur de l'écart type. La division par n-1 donne un estimateur non biaisé. Cette qualité le fait souvent préférer d'autant plus qu'il donne , pour n petit, une estimation plus grande que si on avait divisé par n donc prudente. C'est de là que vient la proposition du logiciel Excel évoquée en II.
L'estimation peut être conduite de manière non paramétrique. Cette approche s'est développée avec la généralisation des ordinateurs. Elle ne demande pas hypothèse sur la forme du modèle qui régit la variable (ou seulement des hypothèses très générales) mais nécessite plus de calculs aussi bien pour obtenir le résultat qui se présente sous la forme d'un graphique que pour l'utiliser.
L'autre grand domaine de l'inférence statistique est celui des tests. Il consiste toujours à se demander si une valeur calculée à partir des données observées est ou non une réalisation anormalement grande ou anormalement petite d'une variable aléatoire dont cette valeur est une réalisation. L'idée sous-jacente est que les probabilités des valeurs anormalement grandes ou anormalement petites sont elles mêmes très petites et qu'un événement de petite probabilité ne doit pas se produire dans une expérience unique, celle d'observer la réalisation d'un seul échantillon. Sur ce sujet aussi , la littérature est foisonnante. Restons proche de notre exemple. La moyenne de la variable paume pour les quinze femmes est 76,6mm. Elle est de 87,1mm chez les trente cinq hommes. Peut–on considérer que cette différence est compatible avec l'hypothèse de l'égalité des moyennes dans les deux populations ? Autrement dit, peut–on considérer que 76,6mm et 87,1mm sont deux fluctuations possibles pour une même variable aléatoire ?
Pour être légitime, le test de comparaison des moyennes doit être précédé d'un test de comparaison des variances. Le tableau 4 fourni par le logiciel Excel montre que l'hypothèse de l'égalité des variances n'est pas rejetée : la variable F objet du test dépasse la valeur trouvée (1,227) avec une probabilité de 0,35 . La valeur trouvée (1,227) n'est donc pas anormalement grande.
Au contraire, dans le tableau 5, la variable T, objet du test, n'est en deçà de la valeur trouvée (-7,26) qu'avec une probabilité extrêmement faible (1,47x10-9). Cette valeur (-7,26) est donc anormalement petite ; elle conduit à rejeter l'hypothèse de l'égalité des moyennes dans la population des femmes et des hommes.

Notons pour terminer qu'un simulateur de la loi de la variable F ou de la variable T, permettrait de construire un histogrammes des valeurs prises par ces variables dans un grand nombre d'appels du simulateur. On pourrait alors situer les valeurs observées dans ces histogrammes et conduire les tests de façon approchée. Cette remarque introduit naturellement certaines pratiques de tests pour des quantités dont on ne connaît pas la forme analytique de la distribution : la forme est approchée par simulation et la valeur observée est située dans cette distribution.
De nombreuses méthodes statistiques (tests de permutation, Jackknife, Bootstrap) utilisent aujourd'hui l'ordinateur pour obtenir par des calculs nombreux ce qui ne peut être atteint analytiquement.

V)Apprentissage et affectation

Supposons maintenant qu'on nous donne une mesure de la variable " paume " et qu'on nous demande sur la base de cette mesure de dire si l'individu sur lequel elle a été faite est un homme. Nous sommes dans une démarche assez fréquente d'étude statistique: les données que nous avons décrites jusqu'ici constituent une échantillon d'apprentissage. On savait pour chaque individu s'il était homme ou femme et on connaissait les mesures faites sur lui. Connaissant maintenant une mesure, on veut savoir si l'individu sur lequel elle a été faite peut être rangé dans le groupe des hommes.
Décidons de ne pas affecter au groupe " homme " un individu dont la mesure appartiendrait à la classe 3 des histogrammes. Ceci est justifié par le fait que l'histogramme issu de l'apprentissage pour le groupe " homme " donne à la classe 3 une fréquence de 3/35 inférieure à 10%. Ce choix fait courir un risque, dit de première espèce, qui est de rejeter le groupe " homme " alors qu'il ne devrait pas l'être et ce risque est égal à la fréquence avec laquelle les hommes sont tombés dans la classe 3, c'est à dire 3/35. Bien sûr, dans des applications réelles, on prend des probabilités inférieures à 10% souvent dictées par le coût économique ou social du rejet à tord.
Inversement, affectons au groupe " homme " un individu dont la mesure appartient à la classe 4 puisque dans l'histogramme issu de l'apprentissage, le groupe " homme " a donné à cette classe la fréquence 11/35 supérieure à 10%. Ce faisant, nous courrons le risque, dit de seconde espèce , de considérer comme issue d'un homme, une mesure qui est en fait issue d'une femme puisque le groupe " femme " a donné dans l'apprentissage la fréquence 3/15 à cette classe. Il n'est pas toujours facile d'apprécier ce risque dit de seconde espèce parce que l'hypothèse alternative , le groupe " femme ", n'est pas toujours aussi simple que dans cet exemple. L'étude de la puissance des tests, c'est à dire de leur capacité à minimiser le risque de seconde espèce est un chapitre important et difficile de la Statistique.

Références : Les livres d'introduction à la Statistique sont nombreux. On trouve aussi aujourd'hui des didacticiels . Nous ne donnerons ici qu'une référence pour l'un et l'autre de ces moyens d'en savoir plus :
Gilbert Saporta : Probabilités, Analyse des Données et Statistique, Edition Technip, 1990.
Collectif : StatNet, http://www.agro-montpellier.fr/cnam-lr/statnet, 2002

Tableau 1

paume	doigt	sexe	paume	doigt	sexe	paume	doigt	sexe	paume	doigt	sexe
80	80	F	94	89	M	72	87	F	87	84	M
75	78	F	88	84	M	81	88	M	83	83	M
71	88	F	87	85	M	97	95	M	85	86	M
69	67	F	90	95	M	77	80	M	93	90	M
75	75	F	94	95	M	91	85	M	94	76	M
75	74	F	90	84	M	89	83	M	91	90	M
77	85	F	85	79	M	90	83	M	79	75	M
78	85	F	90	84	M	81	80	M	84	84	M
76	70	F	85	83	M	80	76	M			
84	83	F	84	91	M	85	80	M			
78	85	F	93	87	M	87	78	M			
74	75	F	82	88	M	82	76	M			
83	85	F	90	83	M	86	94	M			
82	81	F	89	83	M	86	83	M			

Tableau 2

Classes	*Fréquence pour les femmes*
70	1
75	6
80	5
85	3
90	0
95	0
ou plus...	0

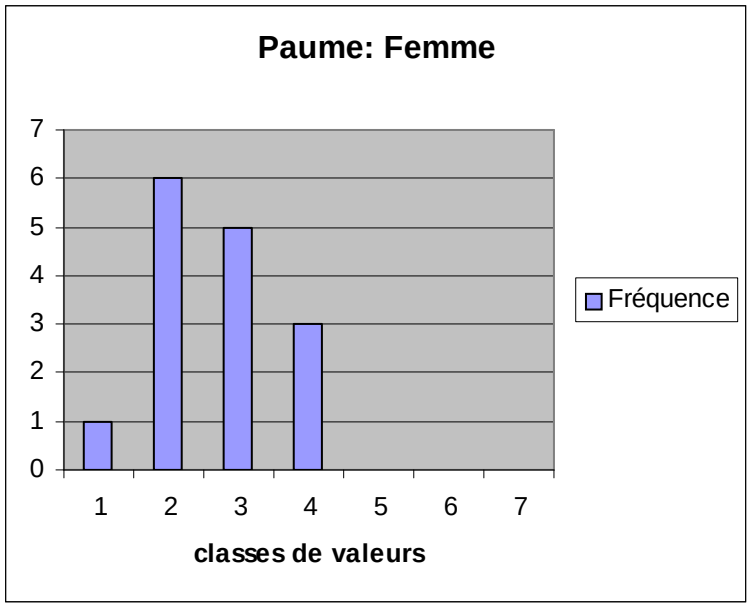


Tableau 3

Classes	*Fréquence pour les hommes*
70	0
75	0
80	3
85	11
90	13
95	7
ou plus...	1

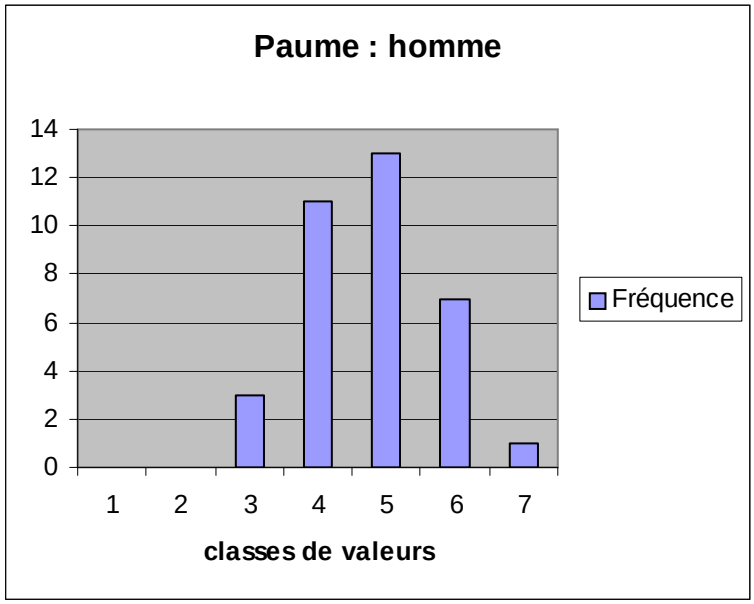


Tableau 4

Test d'égalité des variances (F-Test)		
paume	*homme*	*femme*
Moyenne	87,1142857	76,6
Variance	23,2806723	18,9714286
Observations	35	15
Degré de liberté	34	14
F	1,22714387	
P(F<=f) unilatéral	0,35150245	
Valeur critique pour F (unilatéral)	2,28869368	

Tableau 5

Test d'égalité des espérances: deux observations de variances égales		
paume	*femme*	*homme*
Moyenne	76,6	87,1142857
Variance	18,9714286	23,2806723
Observations	15	35
Variance pondérée	22,0238095	
Différence hypothétique des moyennes	0	
Degré de liberté	48	
Statistique t	-7,25985928	
P(T<=t) unilatéral	1,4708E-09	
Valeur critique de t (unilatéral)	1,67722419	
P(T<=t) bilatéral	2,9415E-09	
Valeur critique de t (bilatéral)	2,01063358	

Postface

Yves Escoufier
Laboratoire de Probabilités et Statistiques
Université Montpellier 2
yves.escoufier@univ-montp2.fr

La mise en forme de cette annexe au rapport sur la Statistique s'achève alors que l'Institut international de statistique (IIS) annonce le lancement d'un grand projet " d'alphabétisation statistique ", connu sous l'appellation anglaise d' International Statistical Literacy Project (ISLP).
La première décision prise dans le cadre de ce projet a été d'inclure un plus grand nombre de termes concernant l'enseignement de la Statistique dans le dictionnaire des termes statistiques de l'Institut international. Ce dictionnaire multilingue est accessible par internet à l'adresse : http://www.cbs.nl/isi en cliquant sur la rubrique ISI Glossary of Statistical terms. Il suffit de repérer un mot du vocabulaire statistique dans une des dix neuf langues connues par le dictionnaire pour avoir sa traduction dans les dix huit autres.

La seconde décision concerne la création d'un site web destiné à apporter des éléments d'information à toutes les personnes concernées par la Statistique. Dans le cadre de cette postface , il n'est pas sans intérêt de relever les objectifs de quelques unes des pages de ce site :

- page B : pour les enseignants de l'école élémentaire, des supports à utiliser dans leurs classes
- page C : pour les enseignants de l'école secondaire, des supports à utiliser dans leurs classes
- page D : pour les enseignants de l'école élémentaire, des supports à utiliser pour améliorer leur connaissance et leur enseignement de la Statistique et des Probabilités
- page E : pour les enseignants de l'école secondaire, des supports à utiliser pour améliorer leur connaissance et leur enseignement de la Statistique et des Probabilités
- page F : des supports pour les enseignants intervenants dans la formation initiale et continue des enseignants.

D'autres pages visent d'autres cibles : les statisticiens des offices gouvernementaux ; les journalistes, les adultes qui veulent apprendre par eux mêmes les bases de la Statistique…
Ce site est accessible à l'adresse suivante : http://course1.winona.edu/cblumberg/islplist.htm .
La page C contient déjà un certain nombre de références. Il n'est pas inutile de souligner que comme l'accès au dictionnaire, l'accès à ce site est gratuit. Le but poursuivi par ses initiateurs est de faciliter l'accès de tous aux connaissances statistiques qui les concernent : la communauté internationale des statisticiens se sent une responsabilité à ce sujet et en assume le coût. Pour le moment, les pages de ce site sont exclusivement rédigées en anglais. Des contributions des statisticiens francophones seraient certainement acceptées puisque le français est l'une des langues officielles de l'IIS. Certains feront – ils l'effort nécessaire ?

Revenons à l'annexe que clôt cette postface. Son objectif n'est pas éloigné de celui des pages E et F citées plus haut : apporter des éléments de culture statistique aux enseignants du secondaire. Les programmes de secondes, premières et terminales des établissements secondaires français se sont récemment enrichis de chapitres de Statistique et Probabilités. De nombreux enseignants de mathématiques se sentent démunis devant cette nouvelle tâche à

assurer. La commission de réflexion sur l'enseignement des mathématiques a voulu les aider d'abord par la rédaction de son rapport sur les Statistiques et les Probabilités puis par la mise en chantier de cette annexe au rapport. Le choix qui a été fait ici est double : il voudrait susciter la curiosité des enseignants pour la Statistique en montrant les interventions de cette discipline dans des domaines variés ; il voudrait également contribuer à la formation de ces enseignants en mettant en œuvre sous leurs yeux les objets statistiques qu'ils enseignent : fréquences, caractéristiques numériques, distributions… Les plus curieux pourront même trouver la possibilité d'aller plus loin en se rapportant à la bibliographie volontairement limitée donnée par chaque article ; elle permet d'entrevoir ce qu'est la recherche dans les domaines de la Statistique concernés par l'article.

Au passage, certaines de ces contributions font toucher les problèmes que l'usage de la Statistique peut poser à la société. Le score de risque bancaire qui est le résultat d'un calcul statistique, peut – il à lui seul déterminer la décision d'accorder ou non un prêt à un individu ? Non, rappelle Gilbert Saporta, mais il a fallu que la Commission nationale informatique et liberté (CNIL) se préoccupe du sujet et délibère. En cherchant à déterminer des courbes de références , Jérome Saracco et ses co – auteurs touchent à la séparation d'individus en deux groupes : ceux qui sont " dans la norme " et ceux qui n'y sont pas. Concernant un problème de cosmétique, l'enjeu de cette séparation ne suscite pas de débat de société. En serait il de même s'il s'agissait de santé ou de résultats scolaires ? La décision au sujet des cosmétiques a cependant des conséquences importantes dans le domaine économique pour l'entreprise concernée. Ce même enjeu économique sous-tend l'article de Pascal Schlich sur les tests de différences en analyse sensorielle. Remarquons que l'auteur souligne que les praticiens de l'analyse sensorielle réclament une norme AFNOR ou ISO pour ces tests et que le statisticien défend l'idée qu'en l'état actuel des connaissances , il serait prématuré de vouloir la fixer.

La Statistique est une discipline impliquée dans les débats de la société ; les conclusions des méthodes qu'elle met en œuvre sont des arguments de ces débats. Contribuons à donner au plus grand nombre la capacité de comprendre les méthodes et leurs résultats : c'était le but de cette annexe .