Statistique et progrès : un mariage heureux ou une cohabitation difficile ?

J.J. Claustriaux

Professeur ordinaire à la Faculté universitaire des sciences agronomiques de Gembloux.
Le texte présenté est celui donné comme leçon inaugurale de l'année académique 2006-2007

Préambule

"Pour innover, la pensée doit faire une association improbable, un coup de poésie qui surprend et éveille".

Cette réflexion de CYRULNICK [2001] introduit une succession de projections qui s'enchaînent selon un parcours aléatoirement chaotique, au rythme de plus en plus orageux et qui est soutenu par une partie de l'œuvre du compositeur romantique norvégien Edvard GRIEG [1843-1907], *Peer GYNT, suite n°1, opus 46* [1888], intitulée *In der Halle des Bergkönigs* (Dans le Palais du Roi de la Montagne). Elle est interprétée par le *London Festival Orchestra* sous la direction de Sven BENGSON et elle vise à placer l'auditoire dans une atmosphère un peu confuse et éloignée du milieu extérieur, situation dans laquelle le statisticien praticien se trouve bien souvent.

Les projections illustrent ainsi le lien entre la théorie et la pratique du traitement statistique des données. Pour la plupart, elles sont extraites de divers travaux de recherche-développement menés à la Faculté universitaire des Sciences agronomiques de Gembloux, dont ceux de BIEVELET *et al.* [2005], BROSTAUX [2006], CARLETTI et CLAUSTRIAUX [2005], DAGNELIE [1959], LEBAILLY *et al.* [2005], LE TALLEC [2006], PALM [2006], RENAUD [2006], ROISIN [2003], SGHAIER *et al.* [2004], etc.

Pendant les projections, l'orateur dépose sur la chaire deux entonnoirs, l'un avec la base du cône tournée vers le haut et l'autre vers le bas.

Introduction

Ah la statistique !

Est-ce l'aventure d'un homme qui voyage dans le monde des trolls, comme vient de nous le suggérer Peer GYNT ?

On pourrait facilement l'imaginer.

En effet, lorsque quelqu'un vous demande quel est votre domaine d'activités et que vous répondez la statistique, très souvent sa mine se transforme. Vous pouvez alors observer les deux réalisations d'une variable alternative.

Soit sa mine se referme et l'individu manifeste une certaine inquiétude. Après quelques secondes d'hésitation, vous considérant comme un génie ou un être anormal, les yeux figés, il vous dit : "*matière difficile n'est-ce pas, je n'y ai jamais rien compris* !"

Soit son visage s'illumine et un large sourire, un peu moqueur, apparaît. En moins de temps qu'il ne faut pour le dire, il vous narre cette fois une petite raillerie sur la statistique, qui est souvent de nature légèrement tendancieuse.

Au cours de ces échanges culturels fortuits, certaines histoires sont très significatives sur le niveau de connaissance du domaine par le quidam, un peu comme s'il suffisait d'obtenir son permis de conduire pour être mécanicien ou de pianoter sur le clavier d'un ordinateur pour être informaticien.

Néanmoins, moi aussi, comme mise en bouche, je ne résiste pas à vous raconter une plaisanterie, certes courte et gentille. Peut-être la connaissez- vous ?

Elle met en valeur un paramètre statistique qui à lui seul ne donne jamais totalement confiance à un statisticien.

Un statisticien est une personne qui ayant la tête dans un four et les pieds pris dans la glace proclame qu'en moyenne il se sent bien ! (Benjamin DERECA).

Allons, après ce préambule animé et cette introduction, soyons sérieux quelques instants.

Le corps de la leçon comprendra quatre parties.

Tout d'abord, il est essentiel de placer la statistique dans son contexte global.

Ensuite, il convient de s'interroger sur l'apport éventuel de la statistique en faveur du progrès, pour autant qu'il soit aussi défini.

Comme j'estime que cela ne sert à rien de demeurer dans le cercle étroit de la statistique et des statisticiens, les quelques exemples choisis pour étayer la première partie de la question seront directement en relation avec des thématiques appliquées, issues du domaine des sciences agronomiques et de l'ingénierie biologique.

Par ailleurs, la statistique est une discipline scientifique à part entière; elle n'est pas un simple outil, une chose matérielle, un truc étrange qui digère en un seul clic des données chaotiques, comme d'aucuns la considèrent souvent depuis que cette machine qu'est l'ordinateur personnel, a permis de la mettre davantage en valeur lui offrant, notamment, de nouvelles perspectives pour mieux encore lui permettre de s'exprimer sous son plus bel aspect.

Mais qui se cache donc derrière la statistique pour la mettre ainsi en évidence ?

C'est tout simplement un personnage comme un autre appelé cette fois le statisticien !

Dès lors, avant de conclure, pour répondre à la seconde partie de la question, il importe d'émettre quelques considérations sur les relations entre ce professionnel qui a comme mission essentielle de faire parler les données, et celles et ceux qui le consultent.

La statistique : vous connaissez ?

1° Comme le définissent différents auteurs, notamment DAGNELIE [1998], le mot statistique est dérivé du substantif latin *status* qui signifie état; il possède deux significations distinctes.

Utilisé le plus souvent au pluriel, le terme désigne tout ensemble cohérent de données, comme les statistiques de consommations des ménages. Employé au singulier, il concerne toutes les méthodes qui permettent, de regarder, de rassembler, d'analyser et surtout d'interpréter des données.

Cette distinction ne consiste pas à séparer deux domaines étanches, car, comme le traitement et l'interprétation des données ne peuvent se faire que lorsque celles-ci ont été récoltées, les méthodes précisent les règles en matière de collecte des données pour que celles-ci puissent finalement être correctement interprétées.

Comme déjà signalé, il est dès à présent essentiel d'insister sur le fait que la matière première de la statistique appliquée, ce sont les données.

En principe, elles sont toutes différentes les unes des autres pour un même phénomène ou processus étudié. La variabilité de ces données est donc la richesse de la statistique et la raison de son existence comme celle du statisticien appliqué qui cherche toujours à découvrir un signal dans ce bruit de fond.

2° Par ailleurs, certains diront que la statistique est une activité et d'autres que c'est une science.

Pour les sciences agronomiques, sciences appliquées par essence, où le risque d'erreur est par principe inévitable, considérons la statistique davantage comme une simple discipline scientifique, un outil au service de quiconque souhaite tester des hypothèses, comprendre ou modéliser un phénomène univarié ou multivarié dont il cherche à maîtriser les paramètres pour ensuite en faire usage en termes de décision ou de prévision.

Cependant, n'oublions jamais que :

"*le modèle n'a pas raison, n'est pas exact, ne se trompe pas; il n'y a pas de modèle qui soit faux. Mais, celui qui use de modèles peut se tromper. Il est entièrement responsable du choix de son modèle et des hypothèses qui le supportent*" [LEGAY, 1997].

Pour encore mieux percevoir où se situe la statistique dans le monde de la recherche de nouvelles connaissances, considérons, certes de façon simplifiée, justement cette notion de modèle qui est fondamentale.

D'une part, on trouve des modèles non expérimentaux qui correspondent aux modèles ou lois mathématiques, comme par exemple tout simplement celui que nous utilisons en Belgique ou en Chine pour calculer le volume d'un cône droit :

$$v = (\pi r^2 h)/3.$$

D'autre part, il y a des modèles expérimentaux, c'est-à-dire des modèles qui sont construits au départ d'observations quantitatives ou qualitatives, avec leurs unités de mesures spécifiques.

Certains de ces modèles sont toujours exacts aux erreurs de mesures près. Ils s'appliquent au Canada comme au Bénin. Ce sont les modèles physiques ou modèles déterministes que tout chercheur espère trouver pour devenir un jour célèbre comme Joule avec sa loi :

$$w = r i^2 t.$$

Heureusement, tout espoir de célébrité n'est pas nécessairement perdu pour le chercheur, car il existe enfin d'autres modèles expérimentaux mais qui cette fois sont entachés d'erreurs incontrôlables, imprévisibles et non maîtrisables, erreurs dites aléatoires et rassemblées par le terme variabilité déjà cité; ce sont les modèles statistiques ou stochastiques comme l'équation du rendement d'une culture ou d'un bio-digesteur soumis aux conditions variables de l'espace ou du temps :

$$y = f(x_1, x_2, \ldots, x_p) + \varepsilon.$$

Néanmoins, l'ingénieur ou le chercheur tentera d'en établir une méthode, une règle et même parfois une police ou une norme, tout en associant un certain degré d'incertitude qui le plus souvent le perturbera ou le dérangera.

Rejoignant Edmond et Jules de GONCOURT, je peux ainsi dire :

"la statistique est donc bien la première des sciences inexactes".

3° De nombreuses publications sont consacrées à décrire l'histoire de la statistique [DAGNELIE, 1995; DROESBEKE et TASSI, 1990] ; une leçon ne suffirait pas pour commenter tout l'historique de la discipline qui remonte à la plus haute antiquité et qui de façon structurée débuterait au milieu du dix-huitième siècle [DAGNELIE, 1998].

La plupart des personnalités célèbres en statistique ont vu leur nom associé à de grandes lois statistiques comme Abraham DE MOIVRE [1667-1754], Daniel BERNOUILLI [1700-1782], Pierre Simon DE LAPLACE [1749-1827], Karl Friedrich GAUSS [1777-1855], Siméon Denis POISSON [1781-1840], Karl PEARSON [1857-1936], William Sealy GOSSET [1876-1937], mieux connu sous le pseudonyme de STUDENT, George W. SNEDECOR [1882-1974], etc.

Au niveau belge, je ne peux omettre de citer Lambert Adolphe QUETELET [1796-1874].

Dans le monde de l'agronomie, il faut aussi mettre en valeur Ronald Aylmer FISHER [1890-1962], dont les travaux ont largement favorisé la pénétration de la statistique dans tout le secteur des sciences de la vie.

A ce propos, la prise de conscience de l'importance de la statistique en médecine a fait apparaître une nouvelle terminologie, plus à la mode, même si elle avait déjà été initiée il y a fort longtemps, notamment par QUETELET, à savoir la biostatistique.

Initialement, ce secteur de la statistique concernait davantage l'étude statistique des paramètres et des fonctions de base de l'être humain, en relation avec son environnement, les questions sanitaires et de toxicité, la prévention des maladies, l'organisation des soins de santé, etc. [ARMITAGE et COLTON, 1998].

Actuellement, la biostatistique concerne toutes les applications des méthodes statistiques aux sciences de la vie [RASCH, 1994], remplaçant ainsi le terme biométrie, mot peut-être un peu trop désuet et rétrograde.

4° Si évidemment l'outil mathématique a toujours sous-tendu les travaux en statistique, un autre outil a réellement contribué à leurs essors théoriques et pratiques, à savoir l'informatique, et surtout, plus particulièrement depuis 20 ans, grâce à la pénétration du micro-ordinateur et des logiciels conviviaux associés.

Ainsi, pour des travaux de recherches conséquents, tout comme en astronomie, ce sont parfois plusieurs processeurs reliés à distance qui se partagent un travail de simulations. Ces dernières permettent de trouver des réponses à des questions compliquées que le mathématicien ne peut pas encore résoudre seul aujourd'hui.

5° Tenant compte de l'évolution de la discipline et des techniques associées, il faut constater ce qui suit.

D'une part, dans les écoles universitaires en statistique, la plupart des matières enseignées ou qui font l'objet de recherches concernent toujours les concepts essentiels suivants :

- les techniques d'échantillonnages,
- la planification des expériences,
- les fondements des méthodes statistiques,
- l'analyse de la variance,
- les modèles linéaires et la modélisation,
- les analyses statistiques à plusieurs variables,
- les méthodes non paramétriques et robustes,
- les séries chronologiques,
- l'analyse des données discrètes,
- l'analyse des données longitudinales,
- les statistiques bayesiennes,
- les logiciels statistiques et le calcul statistique sur ordinateur.

D'autre part, on doit se réjouir de la combinaison de ces matières ou de certaines parties d'entre-elles qui ouvrent alors des particularités ou des approfondissements plus sectoriels ciblés et qui sont souvent l'affaire de spécialistes, davantage en relation directe avec un secteur d'activités économiques particulier.

Citons comme exemples :

- la dendrométrie,
- la génétique quantitative,
- l'épidémiologie,
- l'analyse des données de survie,
- les essais cliniques,
- l'analyse des risques,
- la chémométrie,
- le contrôle statistique de la qualité,
- l'analyse statistique des génomes,
- la géostatistique,
- l'analyse statistique des images,
- le data mining,

  le data warehouse.

6° Pour les sportifs que je ne voudrais pas décevoir et à titre d'exemple, je souhaite les informer que la statistique a aussi envahi les pelouses des stades de football et qu'elle s'intéresse, notamment, à la modélisation d'une rencontre autour du ballon rond en mettant au point des estimateurs au sens du maximum de vraisemblance de l'avantage de jouer à domicile [HIROTSU et WRIGHT, 2003]; comme quoi la statistique mène à tout, il suffit d'observer !

Enfin, à la veille des élections, je m'en voudrais de discourir sur l'usage de la statistique dans les sondages, en particulier de vous parler des marges d'erreur relatives, au risque de faire passer ma discipline pour celle du mensonge.

Et le progrès alors ?

Après ce voyage au pays de la statistique, j'espère vous avoir au moins persuadé de sa raison d'être, de son importance et surtout de sa diversité.

Relions maintenant statistique et progrès, ce dernier mot me semblant plus facile à définir.

Issu du mot latin *progressus*, qui signifie action d'avancer, nous le considérons dans le sens de tout ce qui concerne, en particulier le développement intellectuel et technique de l'humanité, laissant ainsi de côté délibérément l'aspect moral de la définition.

Cependant, j'ai tendance à lui associer un autre mot, à savoir l'innovation, pour introduire tout ce qui a de nouveau dans la connaissance ou dans la technique.

Evidemment, en termes de progrès de nouvelles connaissances, j'imagine que la partie précédente de l'exposé vous a déjà convaincus, au moins partiellement, de l'apport de la statistique en la matière.
Dès lors, vous ne serez pas non plus étonnés d'apprendre que des travaux concernent la statistique et l'innovation.

Cette relation semble parfois si importante au point de vue économique qu'en juin dernier, pour sa nouvelle orientation de sa politique de recherche, la Norvège vient de créer 14 centres d'innovations par la recherche, dont un centre intitulé "*Statistics for innovation*" et sa mission sera, notamment, d'encourager et de renforcer les liens entre les aspects quantitatifs de la recherche académique d'excellence et la recherche industrielle.

Un mariage heureux ?

Mais je perçois que vous n'êtes pas encore totalement satisfaits avec ce seul argument pour sanctionner favorablement ce mariage.

Aussi, par quelques exemples, étayons davantage.

1° A la suite des différentes crises, dont celle dite de la vache folle (encéphalopathie spongiforme bovine ou ESB), la protection de la santé animale et humaine s'est renforcée.

En particulier, des outils analytiques furent mis au point pour permettre des analyses rapides des aliments destinés au bétail et toute la spectrométrie de réflexion diffuse dans le proche infrarouge (technique SPIR) s'est développée (FERNANDEZ PIERNA *et al.*, 2004).

Cette méthode d'analyse fournit la composition chimique d'un échantillon exposé à un rayonnement proche infrarouge à partir de son spectre, c'est-à-dire de la mesure de l'absorption du rayonnement par l'échantillon en fonction de la longueur d'onde ou de la fréquence du rayonnement incident.

Parmi les avantages de la technique, il faut citer, notamment, la rapidité, la simplicité de la mesure et surtout l'aspect non destructif de l'échantillon.

Son utilisation dépasse largement le contrôle dans le secteur animal; elle concerne également tout le secteur végétal, ainsi que la détection de constituants illicites.

La méthode requiert cependant la construction de modèles statistiques appelés les équations de calibrage, domaine particulier de la chémométrie, au départ d'un nombre considérable d'échantillons et surtout de données très particulières par échantillon que sont les centaines de longueurs d'ondes.

Dans ce cadre, une méthode mise au point en 1966, mais seulement appliquée vingt ans plus tard, la régression des moindres carrés partiels ou régression PLS (*partial least squares modelling in latent variables*), est largement utilisée. Plusieurs algorithmes ont amélioré et améliorent encore aujourd'hui la technique de calcul.

Succinctement, la modélisation s'effectue en deux étapes. La première étape correspond à une opération de compression des données (matrice **X**), au cours de laquelle des variables latentes (matrice **T**) en nombre limité sont estimées. Au cours de la seconde étape, le modèle est construit en exprimant la variable dépendante (vecteur **y**) en fonction des variables latentes (PREVOT, 2004) :

$$\mathbf{X} = \mathbf{T}\,\mathbf{P}' + \mathbf{E}_1,$$

$$\mathbf{y} = \mathbf{T}\,\mathbf{q} + \mathbf{E}_2.$$

Pour améliorer le calibrage multivarié, c'est-à-dire pour mettre au point des modèles dont l'erreur de prédiction sera de plus en plus faible, en particulier pour des tailles d'échantillons de plus en plus petites, le statisticien étudie même des méthodes non issues de sa boîte magique traditionnelle.

C'est ainsi qu'il s'intéresse, notamment, à des méthodes qui fonctionneraient comme le cerveau, appelées méthodes connexionnistes ou neuronales pour lesquelles rien ne serait possible sans le support de l'ordinateur.

2° Parfois, il est utile de rassembler l'information.

Que ce soient dans des études en relation avec le développement durable, notamment de la biodiversité, pour lesquelles des masses considérables d'informations sont récoltées ou des recherches qui tendent à comprendre notre comportement alimentaire, le statisticien se trouve face à des tableaux de données qu'il va falloir tenter de rassembler et de structurer.

Dans ce type d'études, l'important n'est pas uniquement la compréhension d'une situation à un moment donné; elle concerne davantage l'analyse de l'évolution d'un phénomène dans le temps, sachant que rien ne lie les variables observées *a priori.*

Les résultats de tels travaux sont, seront ou devraient être stratégiques pour des décideurs, en vue de favoriser tel secteur de la production, comme celui de la production agricole en particulier.

Aussi à titre d'exemple, considérons les données issues des enquêtes sur les dépenses des ménages en produits alimentaires, réalisées par l'Institut National de Statistique (INS) au cours des années 1999, 2000, 2001, 2002 et 2004.

Les dépenses de consommations alimentaires sont relatives à 39 variables et elles concernent neuf catégories de produits, à savoir :

- les pains et les céréales,
- la viande,
- les poissons,
- le lait, les fromages et les œufs,
- les huiles et les graisses comestibles,
- les fruits,
- les légumes, les pommes de terre et les autres tubercules,
- le sucre, les sucreries et la confiserie,
- les autres produits alimentaires.

Il est facile de rassembler les données sous la forme de cinq tableaux bidimentionnels qui, assemblés, constituent un tableau à trois entrées dont les axes sont les variables, les ménages et les années.

De façon générale, le tableau tridimentionnel peut être vu comme une juxtaposition de K tableaux ($\mathbf{X}_k$) à deux entrées de I variables et J individus ou années [BIEVELET, 2006].

La question est simple : comment analyser cet ensemble de données ?

Des méthodes ont été adaptées ou inventées, il y a à peine dix ans et je les résume comme suit.

Elles se fondent sur une analyse multivariée bien connue, l'analyse en composantes principales (ACP), technique descriptive permettant d'étudier les relations qui existent entre des variables quantitatives, sans tenir compte, *a priori*, d'une quelconque structure, ni des variable, ni des individus [PALM, 1998].

On y recherche des combinaisons linéaires des variables qui sont en nombre limité et indépendantes les unes des autres; elles sont appelées composantes principales. Le positionnement des variables initiales dans l'espace limité aux composantes principales et les relations de proximité permettent alors au chercheur d'interpréter son tableau de données.

Dès lors, on peut comprendre que dans le cas de tableaux à trois dimensions, différentes nouvelles procédures peuvent être appliquées comme :

- pratiquer une ACP sur chaque tableau séparément et comparer les résultats (analyse séparée), avec les difficultés d'interprétation que cela représente,
- effectuer une ACP sur un tableau moyen résultant de la combinaison de tous les tableaux (analyse groupée), en perdant toute l'information liée, par exemple aux années,
- juxtaposer tous les tableaux pour en créer un seul dont une dimension est liée à la fois aux individus et aux années et l'autre dimension aux variables, pour ensuite appliquer une ACP sur le tableau résultant (analyse mélangée), en perdant toute l'information spécifique aux individus et aux années.

Les nouvelles méthodes cherchent à palier ces inconvénients, comme l'analyse triadique partielle, l'analyse de co-inertie multiple, l'analyse factorielle multiple, etc.

Et c'est là que la statistique appliquée apparaît pour, tout d'abord, évaluer sur des jeux de données réels la qualité des ces méthodes, ensuite, suspecter quelles en seraient les causes des différences éventuelles, celles-ci n'étant pas nécessairement théoriques, mais davantage liées aux conditions complexes dans lesquelles les données ont été récoltées et, enfin, simuler des situations analogues pour déterminer si ces méthodes donnent finalement des résultats semblables ou dans quelles conditions il faut appliquer l'une ou l'autre méthode.

Le travail de sélection ayant été fait, dans certains cas aussi celui d'améliorations, les résultats peuvent alors être diffusés aux praticiens et les données initialement fournies correctement analysées et interprétées.

3° Et si maintenant, nous consommions ?

Notamment pour boire ou manger, comme acheteurs, nous souhaitons que le produit acheté soit conforme à ce que le fabricant nous propose. Quant au fabricant, il a tout intérêt à produire ce qu'il annonce, faute de quoi, tenant compte de la concurrence actuelle ou de la législation normative, tôt ou tard il disparaîtra.

C'est ainsi que dans l'industrie manufacturière, la statistique a trouvé une nouvelle jeunesse en relation avec la démarche de la qualité totale. En particulier, une fois de plus et comme exemple, des outils statistiques, souvent plus simples que ceux trop rapidement exposés préalablement, ont largement contribués à développer la maîtrise statistique des procédés (MSP). Aujourd'hui, ses principes sont appliqués à de nombreux secteurs d'activités, y compris à celui des services et des laboratoires.

Le but est simple : utiliser des paramètres statistiques pour surveiller la fabrication en série, afin de pouvoir intervenir avec des facteurs correctifs lorsqu'il y a dérive.

A votre santé !

L'embouteillage d'une bière, dont les qualités intrinsèques ne sont pas en cause, comme celle que nous aurons bientôt l'occasion de déguster, si je termine, doit être parfait; même au plan financier, le brasseur n'a pas intérêt à dépasser le volume annoncé même si vous le désiriez et, dès lors, de commun accord, vous souhaitez en recevoir au moins pour votre argent.

Les outils statistiques utilisés sont simples. On a repris la moyenne et l'écart-type dont l'origine remonte il y a fort longtemps; mais, on en a fait bon usage pour concevoir les célèbres cartes de contrôles de SHEWHART, automatisées ou non, vers les années 1930, même si ce n'est que vers les années 1950–1960 qu'elles commencèrent à pénétrer lentement dans les entreprises [PILLET, 1995].

Cette MSP permet :

- d'analyser les fluctuations de la variable aléatoire relative à la caractéristique suivie, par exemple un volume,
- de comprendre la variabilité des observations, afin de discerner si, à un moment donné, il y a présence d'une cause spéciale qui a provoqué ces variations ou non,
- de décider que le processus de fabrication est bien maîtrisé, autrement dit qu'il est sous contrôle, afin d'intervenir dans le cas contraire,
- d'évaluer le niveau de qualité qu'on peut attendre du procédé, aussi nommé capabilité, et ce, pour des conditions de contrôle définies, c'est-à-dire en tenant compte des limites de spécifications ou de tolérances.

Mais attention, ce domaine est parfois bien plus compliqué qu'il n'y paraît.

Pour vous en convaincre, devenez durant quelques instants le responsable de production pour la fabrication de tartes aux fruits.

Votre production sera des tartes constituées, notamment, de pâtes cuites et garnies de fruits; en principe, elles seront toutes identiques alors qu'à l'origine du processus vous allez fabriquer des pâtons qui eux seront frais et donc vous devrez mettre au point un modèle de maîtrise de votre processus qui reliera la tarte cuite et le pâton frais..., pour finalement satisfaire vos clients !

4° Depuis le début de cette leçon, vous vous demandez certainement pourquoi deux entonnoirs sont posés sur la chaire ?

Je vais enfin vous éclairer.

La statistique appliquée est une discipline qui nécessite souvent de communiquer clairement, précisément, de façon répétée et, surtout, en frappant l'imagination pour sensibiliser l'interlocuteur à l'utiliser avec respect et à s'en souvenir.

En préparant l'exposé, mes collègues et moi, nous nous sommes interrogés pour savoir ce que vous garderez en mémoire en quittant cette salle, abstraction faite de l'immense plaisir des diverses rencontres ?

Quel objet, utilisé dans la vie courante, pouvait au mieux symboliser la statistique appliquée ? Quel serait cet ustensile qui nous facilite parfois certains travaux et lui correspondrait, puisque c'est bien là la mission première de la discipline, assister pour résoudre des questions complexes ?

Nous avons pensé que l'entonnoir était cet ustensile.

Positionné avec la base du cône tournée vers le haut, l'entonnoir représente ce que je vous ai déjà exposé, à savoir un outil qui vous aide à rassembler des quantités de données, pour concentrer l'information dans des paramètres, comme le sac qui accueille les grains de froment en passant par la trémie.

Si la base du cône est tournée vers le bas, que symbolise-t-il ?

Les deux derniers exemples que je vais vous raconter vous permettront de l'imaginer et je les résume immédiatement comme suit.

Comment organiser au mieux les unités sur lesquelles je vais effectuer des observations pour éprouver un phénomène, comme évaluer l'intérêt d'un nouveau médicament, d'un nouvel aliment, d'une nouvelle variété de plante, d'une nouvelle race animale, ou encore d'un nouveau bain mousse, sous les conditions que :

d'une part, cela me coûte financièrement le moins possible, à la limite rien, pour trouver la réponse le plus rapidement possible, à la limite

instantanément, ce qui est aussi évidemment absurde,

et, d'autre part, que les résultats, après passage dans le tube de l'entonnoir, puissent être étendus ou dispersés grâce au cône à l'univers tout entier?

5° Nous entrons ainsi de plein pied dans le vaste domaine de la planification expérimentale.

Comme premier exemple, intéressons-nous à l'optimisation d'un milieu de culture complexe pour la production de lipase par une souche de levure [UWAMWEZI, 1996].

Il convient de concevoir une expérimentation qui tienne compte de l'association des constituants du milieu de culture, appelés facteurs étudiés, au nombre de quatre, chacun étant considérés à différents niveaux de concentrations.

La mise en œuvre de l'expérience doit aussi prendre en compte les moyens disponibles afin d'envisager quand même quelques répétitions des combinaisons retenues des facteurs et, surtout, de l'hétérogénéité inhérente à tout domaine dans lequel se déroule une expérimentation. Il n'y a rien de plus variable qu'un espace expérimental déclaré parfaitement contrôlé.

Dans ce cas précis où le nombre total de combinaisons de toutes les sources de variation à considérer peut être de plusieurs dizaines, si pas de plusieurs centaines, les techniques de conception d'expériences factorielles, dont les plans fractionnaires, ont permis des avancées considérables, notamment, pour accélérer le processus de mise au point de la production [BOX et WILSON, 1951; DAGNELIE, 2003; MYERS et MONTGOMERY, 1995].

Certains de ces dispositifs expérimentaux ont été adaptés à des conditions industrielles spécifiques qui tiennent compte, non seulement, du nombre souvent limité d'unités expérimentales disponibles (fours, bio-digesteurs, etc.), mais aussi, de la manière dont il faut ordonner dans le temps la succession des combinaisons des facteurs.

Des adaptations, comme l'approche TAGUSHI ou d'autres avancées qui en sont dérivées [DROESBEKE *et al.*, 1997], sont simples à comprendre dans les principes pour celle ou celui qui veut bien se souvenir qu'il ne faut quand même pas jeter une trop grande quantité de sel de cuisine dans l'eau de cuisson des pommes de terre, le retour en arrière étant bien difficile, n'est-ce pas ? Ainsi, par exemple, un nombre limité de changements d'état peut être fixé pour certains facteurs.

6° Parfois, certains travaux de recherches théoriques en statistique, développés sous la contrainte économique des autorités et pour lesquels le statisticien dénonçait *a priori* l'intérêt, travaux dont l'application initiale a été finalement abandonnée, la raison triomphant, trouvent heureusement quelques années plus tard un nouveau développement et même avec un intérêt financier certain.

Tel est le cas, pour les essais en champs, des dispositifs expérimentaux équilibrés pour le voisinage [AZAÏS *et al.*, 1993].

Considérons une expérimentation destinée à comparer différentes variétés de pois dont les plantes ont la particularité d'être très volubiles. Dès lors, en principe, il faut une taille de parcelle suffisante pour obtenir un ordre de grandeur du rendement représentatif; c'est ici aussi qu'on comprend l'intérêt des bordures.

Mais, par ailleurs, considérons une autorité scientifique qui estime qu'il faut réduire les budgets selon le principe que plus une parcelle est grande et plus elle coûte cher !
Dès lors, le statisticien a la charge de trouver la solution, c'est-à-dire simplement de concevoir des dispositifs expérimentaux dont les parcelles seront les plus petites possibles et sans bordure, sachant que le rendement estimé par la suite sera effectivement et uniquement celui de la variété semée sur la parcelle considérée, sans aucune influence des variétés voisines.

Le dispositif annoncé fut conçu de même que la méthode associée d'analyse des données. En particulier, chaque variété trouve au moins un fois à sa droite et à sa gauche chacune des autres variétés.

Malgré toute la puissance de la théorie et des moyens de calculs, les résultats n'ont jamais été satisfaisants pour des parcelles de trop petites dimensions et ce fut alors un retour vers l'application des bons principes de base de l'expérimentation, à savoir pour un nombre limité et suffisant de répétitions, mettre en place des parcelles suffisamment grandes avec bordures.

Quelques années plus tard, ce dispositif adapté fait son entrée dans les essais sensoriels de

dégustation où de façon équilibrée chaque produit est évalué au moins une fois après le test de chacun des autres produits [CLAUSTRIAUX, 2001].

7° De nombreux autres exemples, comme encore celui qui concerne l'intérêt nutritionnel des acides gras de la famille ω3 [(BURNY, 2006], auraient pu illustrer l'importance d'une planification expérimentale rigoureuse en faveur du progrès des connaissances, au service finalement de l'être humain.

Avant de clôturer cette partie de l'exposé, je souhaite cependant citer DESCARTES [1637] :

les expériences sont d'autant plus nécessaires qu'on est plus avancé en connaissance,

et rappeler si besoin ce qui suit.

Tout d'abord, derrière toute planification expérimentale se cache un modèle statistique d'analyse des données qui imposera tôt ou tard ses conditions d'application ; il vaut donc mieux prévenir que guérir ou mieux encore réfléchir avant, au lieu de soigner après.

Ensuite, même si je considère que la planification expérimentale décrit les concepts fondamentaux de la méthode scientifique de recherche de nouvelles connaissances, il faut parfois accepter que la théorie et la pratique se désaccordent parce que l'expérimentation n'est guère possible pour des phénomènes passés, comme par exemple en géologie ou en paléontologie, lorsqu'un malade est guéri et qu'il n'est plus souffrant ou simplement parce que le chercheur découvre en cours d'expérience un nouveau signal qui le détourne de son chemin initial [CLAUSTRIAUX, 2006].

Enfin, n'oublions pas non plus que tout en défendant la nécessité de recourir à une planification expérimentale rigoureuse pour approcher de nouvelles voies de la connaissance, il ne faut pas ignorer l'importance de la méthode symbolique de recherche qui peut aussi de façon beaucoup moins structurée ouvrir de nouveaux horizons, à condition ensuite de les éprouver, car si on peut de mieux en mieux étudier les mécanismes psychiques de l'amour ou faire l'étude d'une œuvre de GRIEG, on ne peut pas encore dire ce qu'est aimer, ni reproduire les émotions, ni les pensées que Peer GYNT nous inspire.

Une cohabitation difficile ?

1° Le métier de statisticien fait aussi l'objet de travaux scientifiques très sérieux; la littérature à ce sujet est importante [COCKERILL et FRIED, 1991; HAHN et HOERL, 1998; HANOUNE, 1998; HOADLEY et KETTENRING, 1990; HUNT, 2000; LANE *et al.*, 1990, SHETTLE et GADDY, 1998; etc.].

Comme vous l'avez certainement constaté, être statisticien ce n'est pas calculer la probabilité que votre voisin soit assis à côté de vous.

Comme vous l'aurez aussi perçu, être statisticien appliqué pour les sciences agronomiques, c'est être baigné dans l'océan de l'incertitude. Aussi, pour voir un peu plus clair, quiconque ne pourra pas exercer ce beau métier sans certaines connaissances en mathématique, sans maîtriser un peu l'informatique, sans avoir une certaine attirance pour la logique et l'organisation, sans être communicateur et surtout sans avoir consacré un temps suffisant à l'étude, notamment, de la biologie et des sciences du milieu. Le statisticien appliqué doit donc avoir une culture générale significative de son domaine et de celui des autres.

2° Mais, il faut aussi le dire, le statisticien appliqué doit en plus être doué d'un certain degré de diplomatie. En effet, souvent, aussi longtemps que la confiance réciproque avec le partenaire qui le consulte n'est pas établie, il est considéré comme un voyeur et un voyant; il dérange.

D'une part, il est un voyeur parce que celui qui le consulte doit souvent lui avouer que ce qu'il a fait n'est pas tout à fait licite ou qu'il n'a pas bien défini ses hypothèses *a priori*; mais il est trop tard; alors que faire ?

Il faut donc que la confiance s'établisse entre le consulté et le consultant, et cela prend du temps.

Malgré ma longue expérience en traitement des données, jamais je n'ai réussi à réduire à moins d'une heure la première consultation. Il faut écouter, comprendre le problème, le répéter tel qu'il vous a été expliqué pour enfin vous entendre dire que ce n'est pas tout à fait conforme à ce qui c'est passé. Il faut encore mettre au point le protocole d'analyse, accompagner et surtout souvent rassurer.
Tiens, j'ai déjà vu cela ailleurs !

Et aujourd'hui, grâce aux procédures sophistiquées des logiciels, souvent inadéquatement

utilisées ou utilisées par le jeu des essais et erreurs, les questions, qui sont posées, sont de plus en plus complexes et les réponses non immédiates. Dès lors, le statisticien cherche aussi et il apprend beaucoup grâce à ces outils ; il faut oser l'avouer.

Dans la majorité des cas, les résultats de longues conversations et des séances de calculs se terminent davantage par des conseils en termes de structuration et d'organisation des intentions de l'interlocuteur. Il faut constater que le statisticien devient ainsi l'assistant social du chercheur pour les aspects quantitatifs uniquement, entendons-nous !

Bien souvent encore lors des consultations, travaillant dans une institution qui était une célèbre abbaye et me confondant avec un moine franciscain, et non bénédictin, du XIV^{ème} siècle, je dois faire des choix méthodologiques et appliquer simplement le principe du rasoir d'Occam [THORBURN, 1918] :

quand on a deux théories en compétition qui permettent de prédire exactement les mêmes choses, celle qui est la plus simple est la meilleure.

D'autre part, le statisticien est aussi un voyant, parce qu'il est capable de calculer des risques et il sait ou peut savoir que poursuivre telle ou telle recherche dans certaines directions n'aboutira qu'à bien peu de résultats tangibles pour le chercheur. Mais, faut-il encore qu'il soit consulté *a priori*, surtout pour la mise en place de nouvelles expériences ou de nouveaux échantillonnages !

Qu'ils sont heureux ces collègues qui travaillent dans le monde de la recherche pharmaceutique humaine où le statisticien intervient obligatoirement dès le départ pour le choix du nombre de données à récolter.

Notez que ne pas consulter systématiquement un statisticien, c'est aussi une question de culture scientifique qui est inconcevable Outre Manche et aux Etats-Unis, tout domaine confondu. Ah si le bon sens imposait cela dans la mise en œuvre des futures écoles doctorales !

3° La statistique appliquée est-elle un bon placement ?

Au niveau du progrès des connaissances et des techniques, la réponse est inconditionnellement affirmative, même si la statistique est toujours une discipline de l'ombre.

En termes de financements dans le secteur public pour des travaux de recherches en statistique, comme en matière de consultations statistiques, il est difficile de réserver de façon privative les gains de l'innovation, strictement à la statistique. Seulement, des collaborations associant la statistique et une autre discipline sont intéressantes. Hélas, rares sont les collègues qui spontanément vous associent alors que la contribution du statisticien était pourtant essentielle à leurs travaux. L'espérance mathématique d'une simple citation dans des remerciements est aussi bien faible. Que ceux qui néanmoins y pensent en soient remerciés !

Il en est tout autrement dans le secteur privé, surtout si celui-ci développe une recherche interne. C'est édifiant ! La profession de statisticien n'est-elle pas aussi parmi celles qui sont les plus citées aux Etats-Unis ? D'ailleurs, si vous vous intéressez aux placements financiers, je vais vous suggérer une stratégie. Avant de placer vos avoirs, renseignez-vous pour savoir si l'entreprise occupe des statisticiens et si oui, vendez dès le moment où vous apprenez certaines restructurations qui les concernent.

En guise de conclusion

1° En matière de progrès, la statistique seule ne peut rien. Mais, toute recherche efficace et efficiente en sciences de la vie est possible grâce à la statistique.

Elle continuera à s'imposer, pas uniquement là où on observe des hommes et des animaux, notamment pour des raisons d'éthique et de bien-être.

2° La statistique cela s'apprend; elle doit se pratiquer pour être utile.

Il ne suffit pas de compter pour faire sa statistique, comme Monsieur JOURDAIN sa prose

C'est pourquoi, je souhaiterais que la statistique soit encore mieux connue, qu'elle soit plus judicieusement utilisée, que si elle pénètre dans la formation, l'enseignement secondaire en particulier, ce ne soit pas par ses principes théoriques qui découragent et ne permettent pas de l'apprécier, mais uniquement en favorisant l'observation de données et leur interprétation par les étudiants eux-mêmes : il y a tant de chose à regarder et à mesurer.

3° Par ailleurs, jamais, la statistique ou un terme associé à la discipline ne doit se transformer en vérité absolue. Même si la statistique est un phare guidant un chemin dans le brouillard du progrès, même si la logique en est une de ses composantes, n'oublions pas que cette logique est le pire ennemi de la vérité, car elle est incapable de voir ses propres erreurs.

4° Enfin, imaginez-vous encore un seul instant que nous vivions sans cette discipline de l'ombre ? On se tromperait énormément !
Et c'est donc pourquoi, même si ce n'est facile de faire cohabiter la statistique, avec les autres disciplines, c'est sans crainte que j'ose finalement proclamer :

Statistique et progrès : un mariage heureux ***et*** *une cohabitation difficile !*

Remerciements

A l'issue de cette leçon, je tiens à remercier Madame J. AUSTRAET, ma Secrétaire, pour la mise au point du document. Elle débute sa dernière année d'une activité professionnelle bien remplie. C'est pourquoi, je lui dédie cette leçon.

Je remercie aussi Monsieur G. DECRAEMER pour sa contribution aux illustrations de la leçon; il n'a pas hésité à mettre en valeur ses qualités artistiques.

Mes remerciements s'adressent également à tous mes collaborateurs d'hier et d'aujourd'hui, grâce auxquels j'ai pu largement illustrer et compléter mes réflexions, même s'il ne m'a pas été possible de mettre en évidence tous leurs travaux.

Monsieur le Professeur P. DAGNELIE doit aussi être remercié; sans la confiance que ce célèbre statisticien, encore plus rigoureux que moi, m'a témoignée, et ses qualités de controverses, je n'aurais pas pu présenter cette leçon. J'ai aussi une pensée à l'égard du Professeur J. TEGHEM dont l'art d'enseigner une matière ardue, la mathématique, n'est certainement pas étranger à tout ce qui a été mis en œuvre pour la préparation orale de cette leçon.

Enfin, je ne peux pas terminer ces remerciements sans penser à toutes celles et à tous ceux, notamment les étudiants, qui ont enrichi le statisticien praticien en n'hésitant pas à le consulter et à lui faire part de leurs difficultés ou inquiétudes dans le traitement de leurs riches données; merci pour cet enrichissement et les nouvelles pistes qu'ils ont ainsi ouvertes au profit de la science et de la discipline statistique en particulier.

Références

ARMITAGE P., COLTON T. [1998]. Encyclopedia of biostatistics. Chichester, Wiley. 6 vol.

AZAÏS J.M., BAILEY R.A., MONOD H. [1993]. A catalogue of efficient neighbour-designs with border plots. *Biometrics* 49, 1252-1261.

BIEVELET C. [2006]. *Méthodes d'analyse des tableaux à trois entrées*. Gembloux, Faculté universitaire des Sciences agronomiques, 149 p.

BIEVELET C., PALM R., CLAUSTRIAUX J.J. [2005]. Observatory of food consumption in Walloon Region (Belgium). *In : Agrarian Prospects XIV*, Prague (République tchèque), 829-833.

BOX G.E.P., Wilson K.B. [1951]. In the experimental attainment of optimum conditions (with discussion). *J. R. Stat. Soc., Ser. B*, 13 (1), 1-45.

BROSTAUX Y. [2006]. Random forests and decision trees classifiers : effects of data quality on the learning curve. XXII[rd] International Biometric Conference, Montréal (Canada). July 16-22.

BURNY A. [2006]. L'ingénieur se délocalise : du fondamental à l'appliqué et retour, dans le monde du vivant. *J. des Ingénieurs* 100, 17.

CARLETTI I., CLAUSTRIAUX J.J. [2005]. Analyse de la variance ou méthodes de transformation en rangs alignés : quelle méthode utiliser quand des conditions d'application ne sont pas rencontrées ? *Bulletin USAMV-CN*, 62, 1-6.

CLAUSTRIAUX J.J. [2001]. Considérations sur l'analyse statistique de données sensorielles. *Biotechnologie, Agronomie, Société et Environnement*, 5, 3, 155-158.

CLAUSTRIAUX J.J. [2006]. Sciences et Symboles. Libres propos autour des méthodes de recherches. *In :* Workshop International "Diversité des Fabacées fourragères et de leurs symbiotes : applications biotechnologiques, agronomiques et environnementales". Alger (Algérie), 331-334.

COCKERILL R., FRIED B. [1991]. Increasing public awareness of statistics as a science and a profession – reinforcing the message in universities. *Amer. Stat.* 45(3), 147-178.

CYRULNICK B. [2001]. *L'ensorcellement du monde*. Paris, Ed. Odile Jacob, 320 p.

DAGNELIE P. [1959]. Le carré latin magique: technique d'analyse de la variance. *Rev. Agric.* 12(3), 3-12.

DAGNELIE P. [1995]. Statistique, biométrie, agronomie: approche historique. *C.R. Acad. Agric. Fr*, 81(8), 33-39.

DAGNELIE P. [1998]. *Statistique théorique et appliquée. Tome 1 : Statistique descriptive et bases de l'inférence statistique*. Bruxelles, De Boeck et Larcier, 508 p.

DAGNELIE P. [2003].. Principes d'expérimentation. Gembloux, Presses Agronomiques, 397 p.

DROESBEKE J.J., TASSI P; [1990]. *Histoire de la statistique*. Paris, Presses universitaires de France, 128 p.

DROESBEKE J.J., FINE J., SAPORTA G. [1997]. *Plans d'expériences*. Paris, Technip, 509 p.

FERNANDEZ PIERNA J.A., BAETEN V., MICHOTTE RENIER A., COGDILL R.P., DARDENNE P. [2004]. Combination of support vector machines (SVM) and near infrared (NIR) imaging spectroscopy for the detection of meat and bone meal (MBM) in compound feeds. *J. Chemom. 18*, 1-9.

HAHN G., HOERL R. [1998]. Key challenges for statisticians in business and industry. *Technometrics* 40(3), 195-200.

HANOUNE N. [1998]. Trois tables rondes sur la formation des statisticiens et les métiers de la statistique. *J. Soc. Stat. Paris* 39(4), 45-53.

HIROTSU N., WRIGHT M. [2003]. An evaluation of characteristics of teams in association football by using a Markov process model. *Statistician* 52 (4), 591-602.

HOADLEY A.B., KETTENRING J.R. [1990]. Communications between statisticians and engineers/physical scientists (with discussion). *Technometrics* 32(3), 243-274.

HUNT L. [2000]. Why do we do it ? Statisticians and the practice of statistics. *Austral. & New Zealand J. Stat.* 42 (1), 43-58.

LANE J., Ray R., GLENNON D. [1990]. Work profiles of research statisticians. *Amer. Stat.* 44(1), 9-13.

LEBAILLY P., CLAUSTRIAUX J.J., DUQUESNE B., PALM R. [2005]. *Les productions animales en Région wallonne, d'une économie de l'offre à une économie de la demande. In:* L'élevage : hier, aujourd'hui, demain. Quelles attentes ? Pour quels enjeux ? Gembloux, Dixième Carrefour des Productions Animales, 5 p

LEGAY J.M. [1997]. *L'expérience et le modèle*. Paris, INRA, 110 p.

LE TALLEC D. [2006]. *Limites de tolérance : méthodes de calcul appliquées au cas d'un échantillonnage aléatoire et simple ou d'un échantillonnage à deux degrés*. Gembloux, Faculté universitaire des Sciences agronomiques, 125 p.

MYERS R.H., MONTGOMERY D.C. [1995]. *Response surface methodology*. New York, Wiley, 700 p.

PALM R. [1998]. L'analyse en composantes principales : principes et applications. *Notes Stat. Inform.* (Gembloux) 98/2, 31 p.

PALM R. [2006]. Etude des séries chronologiques par les méthodes de lissage. *Notes Stat. Inform.* (Gembloux), 22 p.

PILLET M. [1995]. *Appliquer la maîtrise statistique des procédés* MSP/SPC. Paris, Editions d'Organisation, 336 p.

PRÉVOT H. [2004]. *Comparaison de méthodes statistiques et neuronales pour l'établissement d'équations de calibrage en spectrométrie de réflexion diffuse dans le proche infrarouge*. Gembloux, Faculté universitaire des Sciences agronomiques, 382 p.

RASCH D., TIKU M.L., SUMPF D. [1994]. *Dictionary of biometry*. Amsterdam, Elsevier, 887 p.

RENAUD F. [2006]. *Cartographie et méthodes d'interpolation spatiale*. Gembloux, Faculté universitaire des Sciences agronomiques, 103 p.

ROISIN C. [2003]. *Quantification de l'hétérogénéité structurale des sols agricoles à partir de données pénétrométriques*. Gembloux, Faculté universitaire des Sciences agronomiques de Gembloux, 257 p.

SGHAIER T., CLAUSTRIAUX J.J., BEJI M.A. [2004]. Intérêt des modèles des plus proches voisins pour le contrôle de l'hétérogénéité spatiale : application à un essai de provenances de pin d'Alep (*Pinus halepensis* Mill.) en Tunisie. *Rev. INAT*, 19, 2, 5-22.

SHETTLE C., GADDY C. [1998]. The labor market for statisticians and other scientists. *Amer. Stat.* 52(4), 295-302.

THORBURN W.M. [1918]. The myth of Occam's razor. *Mind* 27, 345-353.

UWAMWEZI M.-C. [1996]. *Dispositifs expérimentaux utilisés en biotechnologiqe : étude bibliographique et application*. Gembloux, Faculté universitaire des Sciences agronomiques, 56 p. + annexes.