Statistique : Introduction

Résumé

Introduction à la Statistique et ses méthodes. Contexte et objectifs (descriptif, explicatif, prédictif) d'une analyse statistique ; les compétences nécessaires.

Ce cours est structuré en deux niveaux principaux et quelques grands thèmes :

- *L : Description et inférences statistiques élémentaires*
- *M1 : Exploration multivariée*
- *M1 : Inférence statistique*
- *M1 : Modèle linéaire et linéaire général*
- *M2 : Modèle linéaire, modèle mixte*
- *M2 : Apprentissage et modélisation*

Réflexions autour de : Statistique et Déontologie scientifique

1 Le métier de statisticien

Le développement continu des moyens informatiques de saisie, de stockage (bases de données) et de calcul permet la production, la gestion, le traitement et l'analyse d'ensembles de données de plus en plus volumineux. Par exemple, les 600 Mb de données produites en une dizaine d'heures par l'un des séquenceurs actuels représentent l'équivalent de la production mondiale déposée dans GenBank entre 1982 et 1996. Les séquenceurs arrivant sur le marché en 2010 produisent en 5 jours 200Gb par traitement. Le perfectionnement des interfaces graphiques offre aux utilisateurs, statisticiens ou non, des possibilités de mise en œuvre très simples avec des outils logiciels de plus en plus "conviviaux". Cette évolution, ainsi que la popularisation de nouvelles méthodes algorithmiques (réseaux de neurones, support vector machine, agrégation de modèles...) et outils graphiques, conduisent au développement et à la commercialisation de logiciels généraux, ou spécifiques à des métiers, qui intègrent un sous-ensemble de méthodes statistiques et algorithmiques plus ou moins exhaustif.

Une question émerge alors de façon très présente ; elle est fondamentale pour l'emploi et les débouchés des étudiants, la gestion des ressources humaines et les investissements économiques des entreprises ou encore les stratégies scientifiques des laboratoires de recherche.

> *Quelles sont les compétences nécessaires à la mise en œuvre de tels logiciels pour analyser, modéliser, interpréter des corpus de données de plus en plus complexes et volumineux produits par une entreprise ou un laboratoire ?*

Les enjeux sont en effet majeurs ; les résultats influent directement sur les prises de décision du *management* ou la validation de résultats scientifiques et leur valorisation par des publications.

2 Terminologie

Le travail du statisticien est d'abord un travail de communication avec des représentants d'autres disciplines ou d'autres métiers. Ceci nécessite beaucoup de rigueur et donc de précision dans l'emploi des mots et concepts lorsqu'il s'agit de traduire en phrases intelligibles des résultats numériques ou graphiques. En effet, de ces *interprétations* découleront des prises de décision.

2.1 Statistique, statistiques, statistique

Le mot *statistiques* avec un "s" est apparu au XVIIIème siècle pour désigner des quantités numériques : des *tables* ou *états*, issus de techniques de dénombrement et décrivant les ressources économiques (impôts...), la situation démographique (conscription...), d'un pays. La Statistique est une sous-discipline des Mathématiques qui s'est développée depuis la fin du XIXème siècle notamment à la suite des travaux de l'école anglaise (K. Pearson, W. Gosset (Student), R. Fisher, J. Neyman...). Une *statistique* est une quantité définie par rapport à un modèle (*i.e.* une statistique de test) permettant d'inférer sur son comportement dans une situation expérimentale donnée.

2.2 Statistique descriptive, inférentielle et apprentissage

De manière approximative, il est possible de classer les méthodes statistiques en trois groupes : celui des méthodes descriptives, celui des méthodes

inférentielles et celui récent de l'apprentissage.

- La Statistique descriptive regroupe les méthodes dont l'objectif principal est la *description* des données étudiées ; cette description des données se fait à travers leur *présentation* (la plus synthétique possible), leur *représentation graphique*, et le calcul de *résumés numériques*. Dans cette optique, il n'est pas fait appel à des modèles probabilistes. On notera que les termes de statistique descriptive, *statistique exploratoire* et *analyse des données* sont quasiment synonymes.
- La statistique inférentielle. Ce terme regroupe les méthodes dont l'objectif principal est de préciser un phénomène sur une population globale, à partir de son observation sur une partie restreinte de cette population, l'échantillon. Il s'agit donc d'induire (ou encore d'inférer) du particulier au général avec un objectif principalement *explicatif*. Ce passage ne peut se faire qu'aux moyens de modèles et d'hypothèses probabilistes. Les termes de statistique inférentielle, *statistique mathématique*, et *statistique inductive* sont eux aussi quasiment synonymes.
- L'apprentissage statistique est issu de l'interface entre deux disciplines : *Statistique* et *Machine Learning (apprentissage machine)*. L'objectif est principalement la construction d'un modèle statistique traditionnel ou algorithmique sans nécessairement d'hypothèse probabiliste, en privilégiant la *prévision* d'une variables qualitative (discrimination ou classification supervisée) ou quantitative (régression). Le contexte est souvent celui de données de grandes dimensions avec comme défi majeur le cas où le nombre de variables explicatives p est considérablement plus important que le nombre n d'observations ou taille de l'échantillon dit d'apprentissage.

D'un point de vue méthodologique, la statistique descriptive précède la statistique inférentielle ou l'apprentissage statistique dans une démarche de traitement de données : ces différents aspects de la statistique se complètent bien plus qu'ils ne s'opposent une fois que le ou les objectifs : descriptif, explicatif, prédictif sont explicités.

Le vocabulaire de la Statistique :

Population Ω (ou population statistique) : ensemble (au sens mathématique du terme) concerné par une étude statistique. On parle parfois de *champ de l'étude*.

Individu $\omega \in \Omega$ (ou *unité statistique*) : tout élément de la population.

Échantillon : sous–ensemble de la population sur lequel sont effectivement réalisées les observations.

Taille de l'échantillon n : cardinal du sous-ensemble correspondant.

Enquête (statistique) : opération consistant à observer (ou mesurer, ou questionner...) l'ensemble des individus d'un échantillon.

Recensement : enquête dans laquelle l'échantillon observé est la population tout entière (enquête *exhaustive*).

Sondage : enquête dans laquelle l'échantillon observé est un sous–ensemble strict de la population (enquête *non exhaustive*).

Variable (statistique) : $\Omega \xrightarrow{X} \begin{cases} \mathcal{E} \text{ si qualitative} \\ \mathbb{R} \text{ si quantitative} \end{cases}$ caractéristique (âge, salaire, sexe, glycémie...), définie sur la population et observée sur l'échantillon ; mathématiquement, il s'agit d'une application définie sur l'échantillon. Si la variable est à valeurs dans $\mathbb{R}$ (ou une partie de $\mathbb{R}$, ou un ensemble de parties de $\mathbb{R}$), elle est dite *quantitative* (âge, salaire, taille...) ; sinon elle est dite *qualitative* (sexe, catégorie socioprofessionnelle...). Si les modalités d'une variables qualitatives sont ordonnées (*i.e.* tranches d'âge), elle est dite *qualitative ordinale* et sinon *qualitative nominale*.

Données (statistiques) : ensemble des individus observés (échantillon), des variables considérées, et des observations de ces variables sur ces individus. Elles sont en général présentées sous forme de *tableaux* (individus en lignes et variables en colonnes) et stockées dans un fichier informatique. Lorsqu'un tableau ne comporte que des nombres (valeurs des variables quantitatives ou codes associés aux variables qualitatives), il correspond à la notion mathématique de *matrice*.

3 Démarche du statisticien

Le crédo de l'enseignant de statistique consiste à répéter inlassablement : un statisticien (ou les compétences qu'il représente) doit être associé *préalablement* à une étude, des expérimentations, une enquête... De la qualité du recueil et de l'organisation des données dépendra bien évidemment la *pertinence* des résultats de l'analyse. Plusieurs questions sont préalables :

3.1 Expérimentation

- Quelle est la question biologique, sociologique, épidémiologique à laquelle je veux apporter une réponse ? En particulier, quel est l'objectif (descriptif, explicatif, prédictif ou une combinaison) ?
- Quelle est la population étudiée ?
- Comment planifier des expériences ou des recueils d'informations dans des bases pré-existantes ?
- Quels sont les échantillons ?
- Précision des conditions expérimentales
- Observations et mesures

3.2 Exploration pour un objectif descriptif

Cette étape est de toute façon un préalable à tout autre objectif. Les données recueillies sont elles de qualité suffisante ? Sont-elles bien exemptes de biais ou artefacts expérimentaux ? Leurs grandes structures (groupes, corrélations...) sont-elles en accord avec les connaissances acquises sur le sujet ?

- Valeurs manquantes, erronées ou atypiques,
- Modalités trop rares,
- Distributions "anormales",
- Incohérences, liaisons non linéaires,
- Transformations, imputation, codage...

3.3 Décision pour un objectif explicatif

Telle variable ou tel facteur a-t-il une influence sur la variable d'intérêt ? Le modèle théorique est-il en accord avec les résultats expérimentaux ?

- Explicitation de l'hypothèse statistique répondant à la question biologique,
- Détermination du modèle statistique correspondant,
- Estimation des paramètres du modèle et calcul de la statistique de test,
- Prise de décision : rejet ou acceptation de l'hypothèse.

3.4 Apprentissage pour un objectif prédictif

Un modèle *explicatif* construit dans l'étape précédente peut être un bon candidat comme modèle *prédictif* mais pas nécessairement. Paradoxalement, un modèle "vrai" n'est pas nécessairement un "meilleur" modèle prédictif s'il est trop complexe, pas assez "parcimonieux". Une quantité impressionnante de méthodes ont été développées ces dernières années sans qu'il soit possible de déterminer, *a priori*, celle qui conduira aux meilleures prévisions sur le problème et les données étudiées.

4 Quel logiciel ?

Deux logiciels sont privilégiés : l'un commercial SAS car le plus répandu et le plus demandé dans les offres d'emplois ; l'autre, R, en distribution libre (licence GNU) comme outil de développement des dernières avancées méthodologiques du monde universitaire.

4.1 SAS

Mis à part le module SAS/IML de langage matriciel très peu utilisé, SAS est un logiciel de type "boîte noire" superposant des couches basses, pour lesquelles l'utilisateur écrit des lignes de code dans une syntaxe complexe, et des interfaces graphiques conviviales (SAS/INSIGHT, SAS User Guide, Sas Enterprise Miner...). Sa diffusion est telle qu'il apparaît en situation de quasi monopole dans certaines branches d'activité comme l'industrie pharmaceutique. Paradoxalement, sa complexité et son coût sont des atouts pour l'emploi de statisticiens indispensables à sa bonne utilisation et donc à sa rentabilisation. Son apprentissage est incontournable.

4.2 R

A l'opposé et à l'exception des traitements les plus rudimentaires pilotés par menu, R est avant tout un langage de programmation pour la manipulation des objets du statisticien : vecteurs, matrices, bases de données, liste de résultats, graphiques. D'un point de vue pédagogique, sa mise en œuvre oblige à l'indispensable compréhension des méthodes et de leurs limites. Il force à admettre qu'il ne suffit pas d'obtenir des résultats, il faut leur donner du sens. Rien ne nous semble en effet plus dangereux que des résultats ou des graphiques obtenus à l'aide de quelques clics de mulot dont ni les techniques, ni les options, ni leurs limites ne sont clairement explicitées ou contrôlées par l'utilisateur. Il est par ailleurs risqué de se laisser enfermer par les seules méthodes et options offertes par "un" logiciel. En pratique, le ré-agencement ou la réorganisation de quelques commandes R offrent une combinatoire très ou-

verte de possibilités contrairement à un système clos de menus prédéfinis. Il offre par ailleurs, grâce à de nombreuses librairies facilement accessibles et continuellement mises à jour, un ensemble exhaustif des techniques et de leurs options ainsi que des interfaces à des gestionnaires de bases de données ou des outils spécifiques à certaines disciplines (Biologie). Les limitations de R sont d'une part celles d'un langage interprété : lenteur pour l'exécution de boucles (à éviter) et d'autre part la taille des données car elles sont toutes chargées en mémoire.

4.3 Quel choix ?

En résumé, il est bien et utile de savoir utiliser ces deux types de logiciels et il est important de comprendre que l'apprentissage syntaxique d'un logiciel est indispensable mais secondaire. Une fois les méthodes comprises et appréhendées, il est techniquement facile de passer d'un logiciel à l'autre, leurs fonctionnalités étant structurellement les mêmes. La difficulté principale ne réside pas dans l'obtention de sorties ou résultats mais dans leur *compréhension*.

5 Domaines d'application

Toutes les méthodes et techniques utilisées nécessitent d'être illustrées sur des exemples simples ou "académiques", pour ne pas dire simplistes, afin d'en comprendre les fondements. Néanmoins, leur apprentissage effectif requiert leur utilisation effective sur des jeux de données en vraie grandeur, issus de différents domaines d'applications. Ce n'est qu'à cette condition que peuvent être appréhendées les difficultés de mise en œuvre, les limites, les stratégies d'interprétation mais aussi la grande efficacité de ces outils.

Ils sont tirés des principaux domaines d'application de la Statistique.

5.1 Sciences de la Vie

Depuis les travaux pionniers de Sir Ronald Fisher, les disciplines des Sciences de la Vie ont toujours motivé les développements de la Statistique : modèles de durée de vie, modèles épidémiologiques, dynamique de population... Les techniques de séquençage et les technologies d'instrumentation à haut débit (transcriptomique, protéomique, métabolomique...) viennent renforcer lourdement cette tendance en posant des défis redoutables au statisticien : que faire lorsque les transcriptions (quantités d'ARN messagers) de milliers de gènes (les variables statistiques) sont simultanément observées pour seulement quelques dizaines d'échantillons biologiques ?

La figure : 1 est un exemple original d'emploi de l'analyse canonique (objectif descriptif). Cette méthode permet de mettre en relation deux paquets de variables (gènes et concentrations d'acides gras) observées sur les mêmes individus (souris).

Le jeu de données utilisé provient de l'Unité de Pharmacologie-Toxicologie de l'INRA de Toulouse. Il concerne 40 souris réparties en 2 génotypes (sauvages et génétiquement modifiées : PPARα déficientes) et 5 régimes alimentaires (dha, efad, lin, ref, tsol). Le plan est équilibré complet : quatre souris par combinaison des deux facteurs.

dha régime enrichi en acides gras de la famille Oméga 3 et particulièrement en acide docosahexaénoïque (DHA), à base d'huile de poisson ;

efad (Essential Fatty Acid Deficient) : régime constitué uniquement d'acides gras saturés, à base d'huile de coco hydrogénée ;

lin régime riche en Oméga 3, à base d'huile de lin ;

ref régime dont l'apport en Oméga 6 et en Oméga 3 est adapté des Apports Nutritionnels Conseillés pour la population française, soit sept fois plus d'Oméga 6 que d'Oméga 3 ;

tsol riche en Oméga 6, à base d'huile de tournesol.

Les expressions des gènes ainsi que des concentrations de 21 acides gras sont mesurées au niveau du foie après euthanasie. Ce jeu de données aux problématiques statistiques très riches est très souvent repris tout au long des présentations des différentes méthodes.

5.2 Marketing

La prospection ou *fouille de données (data mining)* est une appellation issue des services marketing spécialisés dans la gestion de la relation client (GRC) (*client relation management* ou CRM). Elle désigne un ensemble de techniques statistiques souvent regroupées dans un logiciel spécialement conçu à cet effet et vendu avec un slogan racoleur (SAS Enterprise Miner) :

> *Comment trouver un diamant dans un tas de charbon sans se salir les mains.*

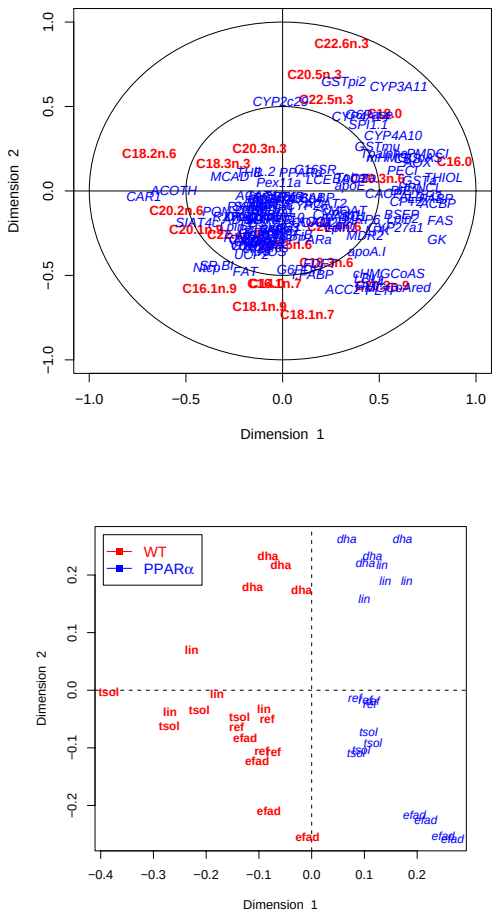

FIGURE 1 – *Souris : premier plan des facteurs canoniques : représentation conjointe des relations gènes et acides gras puis des souris selon le génotype et le régime suivi.*

Les entreprises commerciales du tertiaire (banques, assurances, téléphonie, marketing directe, publipostage, ventes par correspondance...) sont en effet très motivées pour tirer parti et amortir, par une aide à la décision quantifiée, les coûts de stockage des téraoctets que leur service informatique s'emploie à administrer.

Le contexte informationnel de la fouille de données est celui des *data warehouses*. Un entrepôt de données, dont la mise en place est assurée par un gestionnaire de données (data manager), est un ensemble de bases relationnelles extraites des données brutes de l'entreprise et relatives à une problématique.

Chaque banque, assurance... dispose d'un fichier client qui, pour des raisons comptables, enregistre tous leurs mouvements et comportements. Les données anonymes en provenance d'une banque décrivent tous les soldes et produits financiers (emprunt, contrats d'assurance vie...) détenus par les clients ainsi que l'historique mensuel des mouvements, nombre d'opérations, de jours à découvert... La base initiale étudiée comprend 1425 clients décrits par 32 variables explicitées dans une vignette décrivant les données.

Le graphique représenté est un grand classique du marketing bancaire. L'objectif (descriptif) de statistique multidimensionnelle est de construire des classes ou segments de clients homogènes quant à leur comportement bancaire. Une fois les classes construites et l'ensemble des clients affectés, l'agent commercial sait quel langage adopter, quels produits proposer au client qu'il a en face de lui. Après une analyse factorielle des correspondances multiples, les clients caractérisés par leur nouvelles coordonnées sont regroupés en classes dont l'explicitation est facilitée par la représentation des modalités de ces classes dans le plan factoriel de l'analyse des correspondances multiples (figure 2). Un autre objectif (apprentissage) est abordé sur ces mêmes données pour la recherche de scores d'appétences ou d'attrition. Les applications marketing sont très nombreuses (intérêts de certains clients pour des produits financiers, risque pour d'autres clients de changer de fournisseur en téléphonie). Elles le sont également dans les applications financières : risque de défaut de paiement d'un client, de ruine d'une entreprise.

5.3 Industrie

Pour des raisons culturelles et historiques trop longues à développer (culture déterministe des Écoles d'ingénieurs...), la Statistique a une place très mi-

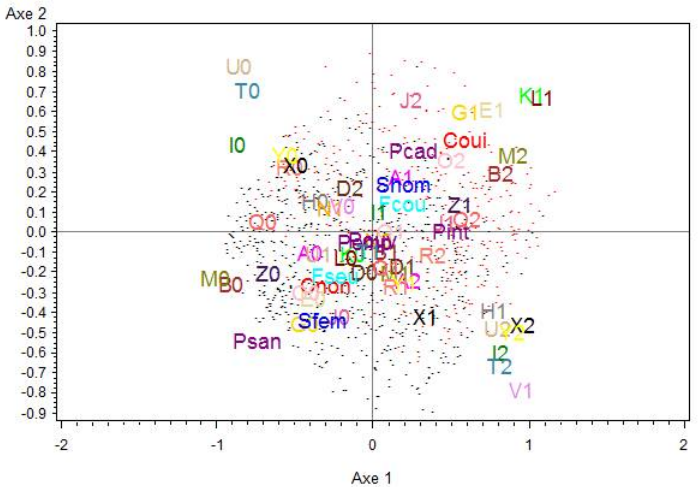

FIGURE 2 – *Banque : représentation des classes de clients, w1 à w5, dans le premier plan factoriel de l'analyse des correspondances multiples*

neures dans l'industrie française sauf en cas d'obligation légale : essais cliniques pour l'autorisation de mise sur le marché des médicaments, contrôle de qualité et fiabilité des matériaux pour la conformité aux normes ISO... La Statistique est ainsi plus vécue comme une contrainte, un contrôle, que comme une aide à la décision. D'autre part, les exemples développés dans le cadre de thèses sont, outre les questions de confidentialité, souvent trop complexes à expliciter pour s'adapter à la simple illustration de ce cours. Néanmoins, il faut être conscient que chacune des techniques abordées, en particulier celles de biostatistique, se transposent directement : durée de vie et fiabilité des matériaux, fouille de données et traçabilité pour la détection de défaillances... Le contexte est souvent techniquement très complexe en terme de modélisation physique mais plus favorable sur le plan statistique, du fait notamment d'un plus grand nombre d'observations que dans le domaine de la santé.

5.4 *Big Data*

Les entreprises industrielles sont actuellement confrontées à la même situation que celles du tertiaire il y a vingt ans : afflux automatique et stockage massif de données. La situation et donc les métiers de la Statistique évoluent considérablement dans ce domaine. Après une période où la question principale est : comment organiser et structurer les matériels et bases de données, la question suivante est : que faire, quelles analyses développées pour les valoriser et aider à la décision ? Prospection numérique dans l'industrie pétrolière, web mining des sites marchands en pleine explosion, utilisation massive des repérages GPS de flottes de véhicules, bâtiments intelligents bardés de capteurs, imagerie 3D... Les applications et problèmes nécessitent en plus, par rapport au data mining maintenant classique, une réflexion approfondie sur les structures de données : fonctions, surfaces, graphes...

6 Quelles compétences ?

Les compétences acquises doivent permettre de répondre avec assurance aux questions suivantes ou alors conduire à une proposition de redéfinition de la problématique envisagée si celle-ci est trop mal engagée.

- Quelle est précisément la question posée ?
- Quelle méthode utiliser avec quelles limites ?
- Comment la mettre en œuvre ?

- Comprendre les sorties du logiciel utilisé.
- Quelle décision ?

Un argument tendancieux, pour ne pas dire fallacieux, est souvent avancé : *il n'est pas besoin d'être mécanicien pour conduire une voiture*. C'est vrai, il n'est pas nécessaire d'être informaticien pour utiliser un ordinateur. En revanche, toute étude statistique nécessite des choix fondamentaux : transformation des données, sélection de variables, choix de méthodes, valeurs des options et paramètres de ces méthodes... qu'il n'est pas prudent de laisser faire, par défaut, au logiciel utilisé. Ces choix ne sont pas anodins et autrement plus difficiles à déterminer que le choix du carburant dans une voiture. Ils doivent être conduits en connaissance de cause par opposition à une stratégie de Shadok (cf. figure 3) qui est un mode d'apprentissage de type "jeux vidéos" : exclusivement par essais — erreurs. Elles est utile, mais pas en toute circonstance, car il ne suffit pas d'obtenir un résultat pour qu'il soit pertinent ou même simplement juste.

FIGURE 3 – *Shadok : devise numéro 1*