
LECTURE 1

STATISTIQUE ET BIOMETRIE: MA VISION

Prof. Kizungu Vumilia Roger

UNIKIN (FACAGRO-BIOLOGIE), UNILU (FACAGRO), UEA (FACAGRO), UCB
(FACAGRO), ISS, ISTA (ENVIRONNEMENT), UPN (FACAGRO-MEDVET)

---GII-GIII-DEA---

Release: 26 décembre 2009

Sommaire

Une enquête.....	3
La définition de la Biométrie	3
La biométrie, un ensemble d'outils pour forger un succès scientifique.....	3
Autant des disciplines, autant de biométries.....	3
Quelles solutions aux grandes barrières de l'apprentissage de la biométrie ?	4
Quid des grands classiques en biométrie ?	4
But du cours en deuxième, troisième graduats et en troisième cycle.....	4
Public Cible	5
Privilégions la question « QUAND utiliser une méthode ? ».....	5
Nécessité des exemples réels et réalistes	5
Niveau des mathématiques et d'informatique	6
Choix d'un logiciel.....	6
Prérequis	6
Évolution du cours.....	7
Challenge ! La Biométrie aux étudiants aspirant aux domaines différents, sans fondement en statistique et en peu de temps.....	7
Focus et contenu	8
Perspectives	8
Soyez indulgents.....	8

Remerciements..... 8

Une enquête

Une enquête auprès des chercheurs révèle que la plupart ont déjà entendu parler de la Biométrie sans savoir exactement de quoi il s'agit. Certains l'associent à **un souvenir douloureux des pratiques calculatoires**. D'autres disent que c'est **discipline à caractère très mathématique** qui aide le chercheur à **analyser statistiquement** les données issues des expérimentations ou des enquêtes. D'autres encore l'associent à **l'interprétation des résultats** obtenus. La majorité des chercheurs ne l'associent pas à la **planification de la collecte de ces données**.

La définition de la Biométrie

Les mathématiques sont vieilles de plusieurs milliers d'années. Elles ont donné naissance, comme discipline, d'abord aux statistiques qui s'en sont démarquées depuis plusieurs siècles et à la Biométrie depuis quelques années.

La définition aujourd'hui acceptée est que la Biométrie est l'**application des mathématiques** en général et des statistiques en particulier dans les sciences de la vie.

La biométrie, un ensemble d'outils pour forger un succès scientifique

Vous avez aussi appris, et vous êtes probablement d'accord qu'il faut comprendre cette discipline pour réaliser du succès dans votre carrière scientifique. Les éditeurs des revues scientifiques exigent le traitement de données par des procédés statistiques. Pour cause, à chaque pas de la vie, vous devez décider et la théorie de la décision se fonde sur des données quantifiées. Par ailleurs, vous avez déjà été victime de la rumeur sur la difficulté de l'apprentissage des mathématiques et des statistiques. Tout cela vous alarme et vous décourage. Surtout si vous n'avez jamais placé les mathématiques et les statistiques parmi les dix premières choses que vous avez décidées d'apprendre durant votre séjour sur la terre des hommes. Ah !!!!

Mon espoir est qu'après ce cours vous changiez d'avis. Pour y arriver:

1. Premièrement, je me permets de vous dire qu'il **n'est pas écrit pour les mathématiciens, encore moins pour les statisticiens**. Il est écrit pour tout chercheur moyen qui garde les notions générales apprises dans un cours semestriel de mathématique en première année d'université. Il n'y sera pas question de l'algèbre, des matrices ou d'autres considérations abstraites.
2. Deuxièmement, je veux amener la Biométrie vers vous comme **un ensemble d'outils**, et non vous vers la Biométrie. La Biométrie doit concourir à la compréhension de votre discipline de recherche et non le contraire.

Autant des disciplines, autant de biométries

Les utilisateurs traditionnels de la Biométrie sont les biologistes, les généticiens, les agronomes et les médecins. Actuellement, cette liste s'étend à d'autres disciplines liées à la science de la vie. Chaque discipline a ses préoccupations. En fonction de différents problèmes, la Biométrie a pris différentes orientations. Il en existe aujourd'hui une pour

médecins, une pour généticiens, une pour la santé publique etc. Toutes utilisent les mêmes principes et chaque groupe focalise sur les modèles proches de ses préoccupations.

Depuis les années 20, la Biométrie et l'expérimentation agricole se marient plutôt bien. Actuellement, avec la bifurcation de l'agronomie vers les sciences biologiques, ce mariage s'est renforcé d'avantage. Analyser les données d'une expérimentation devient similaire à installer un essai. Les deux pratiques se font dans la même logique qui consiste premièrement à identifier le problème puis à décider sur une action à prendre.

Quelles solutions aux grandes barrières de l'apprentissage de la biométrie ?

Les grands problèmes de la Biométrie résident à la fois dans l'application des théories et des formules mathématiques et dans le volume des calculs issus des données observées. En ce qui concerne les théories mathématiques, mon cours élague les démonstrations et présente les résultats à appliquer. Pour ce qui est du volume de calcul, le développement rapide de l'informatique, spécialement des logiciels gratuits a apporté un souffle nouveau dans l'étude de la Biométrie. Il a augmenté la capacité d'analyser rapidement et avec beaucoup de précision des volumes des données beaucoup plus grands. Les sujets comme l'analyse multivariée, qui jadis étaient traités seulement théoriquement jusque vers les années 80 trouvent aujourd'hui des applications.

Quid des grands classiques en biométrie ?

En écrivant ce cours ma préoccupation a été double. Premièrement, je m'interroge comment ce nouveau texte va-t-il différer de ceux, comme Dagnelie (2003), Gomez (1984), Tomassone (1993) qui par ailleurs sont disponibles et se sont imposés comme classiques. Deuxièmement, quelle est sa contribution dans le domaine de l'ingénierie biologique, agronomique et environnementale.

But du cours en deuxième, troisième graduats et en troisième cycle

Le but global de ce cours est de présenter à l'apprenant un ensemble des lectures qui reprennent **les techniques pour modéliser les phénomènes biologiques**. Nous sommes sans ignorer que l'utilisation des modèles et des méthodes statistiques pour décrire et analyser les données est une pratique qui s'impose dans toutes les disciplines scientifiques.

Ces lectures présentent les modèles et méthodes qui ont fait le succès des chercheurs de par le monde dans leur carrière scientifique. Elles sont exploitées pour confectionner un cours trimestriel en deuxième graduat (30 h Cours, 15 h TP). Elles permettent dans ce cas, d'exposer la statistique descriptive univariée et bivariée, la modélisation des observations ou l'introduction aux notions de probabilité et l'inférence statistique ou les théories d'estimation et décisionnelle. Elles permettent aussi d'élaborer un cours trimestriel en troisième graduat. Pour cela elles contiennent:

- le prolongement de la statistique descriptive jusqu'à l'analyse multivariée,

- l'argumentation sur la nécessité de planifier une expérimentation ou une enquête,
- les principes de base de cette planification,
- l'inférence statistique et
- les modèles courants.

Elles sont exploitées pour conférer les cours modulaire au niveau de troisième cycle où les discussions essentielles tournent autour des alternatives aux modèles courants. Bref, elles présentent les outils nécessaires pour débroussailler **un sujet de thèse déjà au niveau du DEA** (école maternelle de la thèse⁽¹⁾) et pour présenter progressivement les articles pendant la thèse.

Public Cible

Ces lectures s'adressent aux étudiants de fin de premier et deuxième cycle qui s'ouvrent à l'art de présenter les résultats de leurs recherches ainsi qu'aux apprenants de troisième cycle. Elles sont conçues pour les professionnels du domaine de la recherche de tous bords: de la santé, de la biologie, de l'agronomie, de la chimie de l'agroalimentaire, des sciences du comportement et qui s'engagent dans la recherche appliquée. Elles sont bénéfiques aux statisticiens professionnels de l'OFIDA, de l'INSS, de l'OGEFREM et des autres organismes spécialisés de l'État qui utilisent des techniques avancées aux problèmes réels de la vie.

Privilégions la question « QUAND utiliser une méthode ? »

Si un cours de Biométrie doit avoir une valeur pour un ingénieur, un chercheur, alors il doit être plus qu'un recueil de formules ou une collection des méthodes à appliquer étape par étape pour prendre une décision dans un univers aléatoire. Le plus difficile, ce n'est donc pas la réponse à la question « comment utiliser une méthode ». En effet, il suffirait de commencer par l'étape A, de continuer par l'étape B puis C et quand l'étape F est atteinte, la solution est trouvée. Ceci laisserait l'étudiant, le chercheur sans comprendre l'importance des statistiques ou de la Biométrie dans la recherche. Le plus difficile donc c'est de savoir « quand utiliser cette méthode ». Ceci exige une connaissance beaucoup plus profonde de la méthode.

Nécessité des exemples réels et réalistes

Dans les sciences de la vie et pour un chercheur, l'usage des exemples réels, des données tirées d'un contexte connu est un moyen de mieux le connecter à la méthode. La brève expérience d'enseignement me le prouve davantage. Les exercices résolus comme exemples en mathématique, donc sans assise sur terrain, ne suscitent pas beaucoup d'intérêt et ne connectent pas les étudiants à la réalité de leur profession. Les exemples de ce cours trouvent leur inspiration dans ma petite expérience comme Directeur de l'Expérimentation Agricole à l'INERA et dans mes séminaires auprès des agents des organismes spécialisés de l'État. A l'INERA, j'ai eu beaucoup de discussions avec les chercheurs affectés à la sélection variétale, à l'agronomie ou à la défense des

¹ Qui enverrait son fils à l'école maternelle sans penser à la suite ?

cultures. Aussi avec les chercheurs plus proches des paysans ou des décideurs de la territoriale.

Dans la mesure du possible, pour chaque exemple, nous donnons ce que l'expérience tentait de prouver, quels étaient les individus en étude, et quelles sont les mesures qui étaient effectuées sur ces individus.

Niveau des mathématiques et d'informatique

Le texte est très modeste quant aux développements mathématiques. Les théorèmes ne sont pas démontrés. Quand c'est nécessaire comme pour le cas du Théorème Central Limite, la démonstration se fait par simulation sur ordinateur. Les calculs sont faits pour des raisons pédagogiques sur de petits tableaux de données. L'ordinateur, actuellement, est un outil indispensable qui facilite l'analyse ou la représentation graphique des données. Les tableaux aux dimensions réalistes sont traités à l'aide d'un logiciel utilisant les mêmes formules utilisées manuellement.

Choix d'un logiciel

Il en existe plusieurs. Chacun a ses particularités. Souvent ils sont écrits pour un public ciblé et présentent leurs forces et leurs faiblesses. Dans ce texte, nous avons choisi d'utiliser le logiciel R. Pourquoi ?

1. Il est polyvalent : R est **un langage de programmation**. Le chercheur n'est pas limité par les procédures préprogrammées des logiciels à menus. Il est aisé de programmer une nouvelle procédure avec R.
2. Il est interactif : l'analyse de données est une opération interactive. Ceci est un avantage sûr qu'a R par rapport aux logiciels à menu.
3. Il est open-sources : le chercheur peut modifier la procédure à son goût. Il peut télécharger sa nouvelle version à tout moment sur www.r-project.org.
4. Il est populaire : le logiciel SAS est le logiciel commun en général. R est le plus populaire. Il suffit de jeter un coup d'œil sur les journaux scientifiques et s'en convaincre. Il existe beaucoup de liste de discussion sur R.

Les avantages qu'offrent R ont un prix. Il est difficile à apprendre. Le cours reprend les commandes de R pour chaque sortie. Ceci fait que l'utilisateur peut commencer à utiliser R avant de l'apprendre en profondeur. Le site cité ci-dessus contient un nombre impressionnant de références sur R.

Prérequis

Ce cours est accessible à tout chercheur qui, dans son parcours, si lointain soit-il, a eu un cours de mathématique pendant un semestre à raison de 4 h par semaine et reprenant les matières sur le calcul matriciel, les équations différentielles et le calcul intégral. Il est accessible aussi à tout chercheur qui a une connaissance élémentaire en informatique et qui doit savoir saisir les données sur Excel, les organiser, les afficher, les sauvegarder à un endroit précis.

Évolution du cours

Pendant l'année académique 2003-2004, le cours s'intitulait à la Faculté des Sciences Agronomiques de l'UNIKIN, en troisième année, "biométrie et dispositifs expérimentaux". Le cours de statistique existait à cette époque en deuxième année. En troisième année, l'attention portait alors sur les dispositifs expérimentaux. A partir de l'année académique 2004-2005, le nouveau programme est appliqué en deuxième année graduat. Le cours de statistique est éliminé du programme. La troisième année continue avec l'ancien programme. L'année académique 2005-2006, le nouveau programme est appliqué en troisième année graduat. Le cours s'intitule alors « Statistique et Biométrie ». Le dispositif expérimental n'est plus le focus. Il est dilué dans les méthodes pour optimiser une décision statistique. Nous avons jugé bon d'utiliser le Modèle Linéaire Généralisé qui est l'unification des autres modèles. Ce fut un langage des sourds avec les étudiants qui n'ont pas eu des statistiques en deuxième année. En 2006-2007 le cours est présenté sous sa forme actuelle c'est-à-dire un ensemble des modèles pour des situations précises dans la vie. C'est permis puisque la Faculté s'est doté de deux salles informatiques de 10 ordinateurs chacun. Le logiciel R est choisi comme outil de travail. Les anciennes versions de SPSS et GENSTAT sont mises à la disposition des étudiants pour comparaison. Un Syllabus est mis en circulation en remplacement de celui de 2005-2006. Il est réédité en 2007-2008. La version 2008 du syllabus est éditée par CEDESURK qui a soumis le texte à une commission scientifique de son choix. Un des Professeurs de Gembloux écrira même dans son rapport qu'il adoptait sa présentation. Les améliorations sont les suivantes pour cette version 2009:

- poursuite avec les questions sur la vérification des connaissances à la fin de chaque chapitre.
- Réformulation et regroupement des exercices.

En 2011 ce syllabus va muer en manuel qui portera le titre "Biométrie et Modélisation, les mécanismes biologiques, les dispositifs expérimentaux et les modèles associés"

Challenge ! La Biométrie aux étudiants aspirant aux domaines différents, sans fondement en statistique et en peu de temps

Ce cours a été écrit pour un public qui n'a aucune formation en statistique ⁽²⁾. Il devait être confiné sur un trimestre à raison de 4 heures par semaine (45 heures). Une autre dimension est de pouvoir le présenter en une semaine pendant 5 jours sous forme de séminaire ou cours intensif. L'objectif était de focaliser sur les principes et méthodes de statistiques. Ces principes sont présentés sous forme des modèles afin d'épouser un concept d'actualité, à savoir la « modélisation ». Notons que les modèles présentés ici ne sont pas les seuls à exister. Ils couvrent néanmoins un vaste champ d'application aussi bien en phytotechnie, en zootechnie, en économie, en gestion des ressources phytogénétiques et naturelles ou en chimie. Pour raison de gestion de temps, les notions importantes pour la compréhension de certains modèles n'ont pas été approfondies. Les analyses non paramétriques sont évoquées ca et là comme alternatives mais ne sont pas bien décrites. Pour la même raison, l'approche logicielle a été adoptée en lieu et place

² Dans le nouveau programme, le cours de statistique n'existe plus en deuxième année

de l'approche pas à pas des auteurs comme Gomez Kwanchai, Pierre Dagnelie. Nous sommes plutôt de la tendance Richard Tomassone. Les démonstrations des formules ont été élaguées à dessein et les étudiants intéressés sont renvoyés à ces auteurs classiques incontournables.

Focus et contenu

Ce cours est écrit pour les chercheurs qui ont des problématiques clairement exprimées en termes des problèmes scientifiques. Le contenu est le même que les livres peut-être plus détaillés et plus complets. L'originalité de ce texte est la présentation de ce contenu sous forme des modèles présentés en fonction des types des variables. Dans le premier chapitre, il m'a paru nécessaire de placer la Biométrie dans la démarche scientifique, d'expliquer la nécessité de la planification d'une expérimentation ou d'une enquête et enfin de donner les principes qui guident cette planification. Ce chapitre a été motivé par une enquête qui stipulait que la plupart des chercheurs identifiaient la Biométrie à l'analyse et à l'interprétation des données et non à la planification de la collecte de ces données ainsi qu'au choix du modèle à ajuster aux données. Le deuxième chapitre introduit la statistique descriptive ou l'art de décrire, de résumer et de représenter les données. Le faire avec R c'est un regal ! Le troisième chapitre introduit les notions de probabilité à travers la notion de modélisation des observations. Le lecteur est invité dans ce chapitre de passer pour la première fois dans le monde de l'abstraction comme les conçoivent les mathématiciens. C'est dans ce chapitre où l'on demande pour la première fois au chercheur d'imaginer que son expérimentation n'est qu'une réalisation parmi tant d'autres. Il referait la même expérimentation dans les mêmes conditions, il n'aura pas les mêmes résultats. Les mathématiciens ont imaginé de les réaliser un nombre infini de fois. Le quatrième chapitre donne en résumé les étapes pour prendre une décision dans le monde aléatoire décrit au chapitre troisième. Du cinquième chapitre au dernier, les problématiques courantes sont exposées ainsi que les outils pour prendre des décisions.

Perspectives

Il est prévu l'ajout des modèles dynamiques simples et une introduction aux séries temporelles. Il est prévu aussi un recueil d'exercices interactifs sur support électronique.

Soyez indulgents

Mes textes actuels sont en cour d'élaboration. Je sollicite votre indulgence du fait que je le mets en circulation sous cette forme provisoire. Je vous serai très reconnaissant de m'envoyer des remarques à l'adresse kizunguvumilia@yahoo.fr pour son amélioration.

Remerciements

Mes remerciements vont droit au Professeur Kalonji Mbuyi qui m'a toujours encouragé pour écrire un manuel de Biométrie pour chercheurs dans les sciences de la vie. C'est lui qui m'y a fermement exhorté en me montrant qu'il y a un vide sur ce plan en République Démocratique du Congo. Parmi les professeurs, je remercie aussi le professeur Kiatoko Mangeye, passionné de la statistique, qui pendant de longues

discussions, m'a montré l'importance des résolutions manuelles sur des cas réalistes. Ceci permet au chercheur qui fait la même chose sur un grand tableau de données et sur ordinateur de savoir exactement ce qui se fait dans l'ordinateur. Pour lui, un logiciel ou un système « presse-bouton » ne sera plus une boîte noire qui débite des résultats dont on ne connaît pas l'origine. C'est suite à ses conseils qu'entre autre, le logiciel R a été adopté. Celui-ci permet de faire plus vite ce qu'on fait à la calculatrice.

Pour trouver les textes de ce cours aller ouvrir une boîte aux lettres biometrieetinformatique@yahoo.fr avec comme mot de passe rvk2009.

Biométriquement vôtre,

Roger KIZUNGU Vumilia

Dr. En Biométrie

kizunguvumilia@yahoo.fr

