
Lecture 2. Place de la Biométrie dans la démarche scientifique

Prof. Kizungu Vumilia Roger

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

---GII-GIII-DEA---

Release: 24 décembre 2009

Sommaire

| | |
|---|----------|
| La démarche scientifique | 2 |
| La démarche de l'expérimentateur | 3 |
| La démarche de la Biométrie..... | 6 |

Introduction

Très souvent, le chercheur planifie son essai, collecte les données puis enfin va voir le Biométricien pour "analyser les données". Ceci met souvent le Biométricien dans une situation incofortable car le chercheur voudrait qu'il comprenne "aujourd'hui, maintenant et immédiatement", sur le champ ce qui lui a pris du temps pour la formulation. Ceci est lié au fait que beaucoup de cette race de chercheurs, en disparition heureusement, croient encore que le Biométricien est cet exécutant mécanique au bout de la chaîne programmé pour travailler en extrême urgence.

Après l'étude de ce chapitre, mon souhait est que tu saches:

- placer la Biométrie dans le contexte d'une démarche scientifique
- connaître la démarche de la Biométrie
- expliquer comment exprimer son thème de recherche sous forme de tableau de données exploitable par tous les logiciels modernes.

La démarche scientifique

Les pays qui ont fondé leurs décisions sur des observations quantifiées sont très avancés. Ceci procède en une démarche bien définie que j'appelle dans la suite, la démarche scientifique. Ce n'est pas la seule et encore moins la plus prépondérante. Elle a seulement le mérite de ne pas laisser beaucoup de place au sentiment ou au subjectif. Les professionnels qui décident de suivre cette voie sont appelés à écrire des monographies sur base des données existantes dans des rapports, des articles etc. Ils sont appelés à élaborer des rapports à l'issue d'une enquête ou d'une expérimentation. Il y a toute une démarche pour collecter ces données. Enfin ils sont appelés à décrire les données ou à décider. C'est aussi une démarche bien élaborée.

La recherche scientifique est avant tout orientée vers le service à la communauté. Elle implique schématiquement 5 étapes :

- l'accumulation des informations autour d'une observation,
- la formulation et l'émission des hypothèses,
- la vérification de ces hypothèses par une collecte de données des rapports, des articles, par une enquête ou par une expérimentation,
- l'élaboration d'une théorie après plusieurs vérifications et
- la restitution à la communauté ou la communication des résultats par les articles, par la radio etc.

L'Observation est faite par un agriculteur, une institution ou un chercheur qui se trouve devant une question, devant quelque chose qui l'intrigue, qui l'interpelle ou le préoccupe.

A partir des observations, le chercheur fait des supputations, émet des hypothèses, avance des explications'

Il peut éclater sa question en plusieurs questions plus précises dites aussi questions scientifiques. Il monte une série d'expériences dites aussi observations provoquées, en vue de confirmer ou d'infirmer les hypothèses émises.

Remarquons que par cette pratique, le chercheur passe d'une hypothèse à tester à une expérience puis à une nouvelle hypothèse à tester etc. Plusieurs expériences donnent lieu à une accumulation des connaissances qui vont être utilisées par les paysans, par les entreprises sous forme des théories vulgarisées. Le processus semble infini. Le chercheur passe d'un problème à un autre, d'une hypothèse à une autre.

La démarche de l'expérimentateur

L'expérimentateur quant à lui procède en quatre choix :

- celui du matériel d'étude appelé individus dans la suite,
- celui des caractères à mesurer et qui varient d'un individu à un autre aussi appelés variables dans la suite,
- celui de la procédure de la mesure de ces caractères appelée planification de la collecte des données dans la suite et
- celui de la prise de décision dite aussi inférence statistique.

Pendant la planification de l'expérience, les deux premiers choix sont celles du spécialiste en biologie, en agronomie. Les deux derniers exigent souvent la collaboration d'un biométricien (¹).

Les individus, les unités expérimentales ou les sujets de recherche

Toute recherche porte sur des *sujets de recherche* dits aussi *individus* ou *unités expérimentales*. Elle est donc impossible à mener si ces individus ne sont pas identifiés sans ambiguïté. Les *individus* peuvent être des chefs de ménage dans une enquête. Ils peuvent aussi être des champs d'observation, des parcelles d'observation, des cultures, des parties de ces cultures (racines, tiges, feuilles, etc.), des ravageurs sur une partie d'une plante dans un essai agronomique. Ils peuvent être aussi des oiseaux dans une observation en biologie. Ils peuvent être des années ou même des jours dans une étude économique ou agroclimatique.

L'ensemble des individus sur lesquels porte une étude est appelé *population* ou *univers* dans le jargon des statisticiens. Sa définition doit être *précise* de sorte que l'on soit certain que tous les individus auxquels on s'intéresse s'y retrouvent et que lors du recrutement pour l'étude, chacun soit sélectionné sans ambiguïté. Quand une étude porte sur cet ensemble, on l'appelle *recensement*.

Cela n'arrive pas toujours. Il peut donc arriver, pour des *raisons économiques, d'inaccessibilité ou de temps*, que l'étude porte seulement sur une partie des individus d'une population. On dira qu'elle porte sur un *échantillon* (²) de cette population. Dans

¹ Entendre toujours, mathématiciens ou statisticiens des sciences de la vie

² L'échantillon est parfois appelé série statistique. Il y a une nuance : dans la suite, le terme échantillon sera réservé aux expériences menées indépendamment les unes des autres dans des conditions identiques (mesure du poids de 100 graines de haricot issues d'une parcelle expérimentale). La série statistique provient des expériences non interchangeables (comptage quotidien des arthropodes sur une culture). En particulier, si elles proviennent d'instantanés successifs, on parle d'une série chronologique.

ce cas, le nombre d'individus retenus pour l'étude est appelé *taille* ou *effectif de l'échantillon*

Les caractères des individus ou les variables

Dans toute recherche, une fois les individus connus, le chercheur détermine les caractères à mesurer sur ces individus. Les *variables* sont donc les attributs qui permettent de caractériser les individus. Elles varient d'un individu à un autre : d'où leur appellation. Leur choix dépend de l'objectif de l'étude. Ce sont des tailles des personnes, des rendements des champs, des hauteurs des plantes, des profondeurs des sols, des nombres de jours pour observer un phénomène, des sexes, etc.

Les variables ont différentes natures. On en distingue deux ; les *variables qualitatives* ⁽³⁾ et les *variables quantitatives* ⁽⁴⁾.

Les variables quantitatives sont *des chiffres sur lesquels les opérations arithmétiques* ont un sens. Elles sont soit *continues* (mesures) soit *discrètes* (dénombrements, comptages).

Les variables *de dénombrement* ou *de comptage* dites aussi *discrètes* appartiennent à l'ensemble des nombres naturels \mathbb{N} . Elles ne prennent que peu des modalités distinctes. C'est le cas du comptage des parasites sur leurs plantes hôtes, de celui du nombre de vache par exploitation.

Les *variables continues* sont des intervalles de l'ensemble \mathfrak{R} des réels. Toutes les valeurs sont *a priori* différentes. Il suffit que l'instrument de mesure soit très précis. C'est par exemple le rendement d'une parcelle, le poids de 100 graines du soja, la taille d'une plante.

On distingue souvent les caractères discrets des caractères continus mais la frontière entre les deux est moins claire en pratique qu'en théorie. En effet, les différentes techniques statistiques tentent toujours de rendre discret les variables continues. En particulier, quand on regroupe une variable continue en classes, la variable qui en résulte est qualitative. La même variable demeure quantitative mais discrète quand on considère les centres des classes.

Les *variables qualitatives* sont discrètes et ont des valeurs isolées qui sont des caractères énumérables dits aussi des modalités. Elles sont soit *nominales*, soit *ordinales*.

Une variable qualitative nominale est celle où l'on classe les individus tel qu'à ceux qui se trouvent dans la même classe, on attribue la même modalité. Ces variables ont la spécificité de ne pouvoir être ni ordonnées ni ajoutées. C'est le cas d'une variable dont les valeurs sont « oui » ou « non ». C'est aussi le cas des couleurs, des numéros d'un territoire, des variétés d'une culture etc.

Les modalités d'une variable qualitative doivent être :

³ Categorical, qualitative

⁴ Numerical, quantitative

- incompatibles c-à-d un individu ne doit pas appartenir à deux modalités différentes,
- exhaustives c-à-d toutes les situations sont prévues et
- sans ambiguïté c-à-d qu'il n'y a pas d'erreur de classement.

Une *variable qualitative ordinale* est celle qui permet d'affecter des individus à des catégories *dans une relation d'ordre*. Elles *peuvent donc être ordonnées mais pas ajoutées*. Les valeurs numériques d'une variable ordinale n'ont que peu de sens et n'importent que par leur ordre. D'où possibilité de comparaison. C'est le cas du niveau d'Azote (80 kg/ha sur la parcelle 1, 110 kg/ ha dans la parcelle 2). C'est le cas d'une variable dont les valeurs ne seraient "jamais", "rarement", "de temps à autres", "souvent", "tout le temps". On peut les substituer par un code numérique pour exprimer l'ordre seulement.

Si une étude porte sur n individus et p variables, les observations forment un tableau à n lignes et p colonnes où les variables sont représentées par les noms des variables ou les identificateurs ou les *entêtes* ⁽⁵⁾. Si p variables sont en étude simultanément, alors l'analyse est dite multivariée de dimension p .

Si on s'intéresse à l'activité sportive de 237 étudiants à qui l'on distribue les fiches d'enquête qui comportent 10 questions alors cette étude a 237 individus et 10 variables. Les variables sont le sexe (variable qualitative nominale), la longueur de la main active et de la main inactive (variables quantitatives continues), la main active (variable qualitative nominale), le nombre de battement de cœur (variable quantitative discrète), la taille (variable quantitative continue), l'âge (variable quantitative continue).

Si on s'intéresse au rendement d'une culture de maïs et que l'on veut connaître l'effet variété (4 variétés) et l'effet fertilisation (3 niveaux de fertilité), on peut répéter les traitements 4 fois. Cette expérience a $4 \times 3 \times 4 = 48$ individus ou parcelles d'observation. Elle a 4 variables à savoir la variété (qualitative nominale), la fertilisation (variable qualitative ordinale), la répétition (variable qualitative nominale) et le rendement (variable quantitative continue).

La recherche appuyée par la statistique

Pour plusieurs personnes encore aujourd'hui, la recherche appuyée par la statistique se résume aux descriptions numériques des données. Certains se focalisent sur le calcul des moyennes, des taux et des proportions en pourcentage. Pour ces personnes, le calcul constitue la première préoccupation qui surplante même celle de la description des données. La recherche proprement dite est laissée de côté.

La recherche est l'étude des relations entre un ensemble des caractères ou variables mesurés sur des individus dits aussi unités expérimentales ou sujets de recherche. Elle est bien accompagnée des techniques ou méthodes qui font l'objet de ce manuel. L'objectif primordial d'une recherche consiste en l'explication de la variation d'une des variables qui est alors expliquée par les autres. On l'appelle variable expliquée ou réponse. On dit qu'elle est prédite à partir des prédicteurs ou des variables explicatives ou encore des variables indépendantes.

⁵ header en anglais

Dans l'exemple sur les étudiants, une recherche peut porter sur le nombre de battement du cœur que l'on cherche à étudier par le sexe. On se posera la question suivante : le sexe a-t-il un impact sur le pouls de l'athlète ? En d'autres termes, est-ce que la variation observée dans le nombre de battement de cœur est due au sexe ?

Expérimentation, Quasi-expérimentation et simple Observation

Quand les sujets de recherche sont assignés de façon aléatoire aux niveaux des variables, on dit qu'on planifie la recherche et donc qu'on fait une **expérimentation**. L'expérimentation est donc une étude contrôlée qui maximise le pouvoir de l'expérimentateur à isoler les effets observés sur une réponse.

Quand les sujets de recherche sont assignés non aléatoirement aux niveaux des prédicteurs, l'étude est dite **quasi expérimentation**. On y gagne en économisant l'argent mais on y perd dans le pouvoir d'isoler l'effet en étude.

Quand les sujets de recherche ne sont pas affectés explicitement aux niveaux des prédicteurs, on dit qu'on effectue une **simple observation**.

La démarche de la Biométrie

La biométrie fonctionne schématiquement en trois phases (Tomassone 1993):

Phase I : Définition du problème scientifique.

Phase II : Recherche bibliographique sur les Modèles existants et Sélection d'un d'entre eux (M). Il s'agit d'inventorier les modèles disponibles susceptibles d'atteindre l'objectif scientifique fixé. En sélectionner un (M). Cette sélection est liée à l'expérience et à la compétence du biométricien ou du biologiste formé en Biométrie et à sa capacité de traiter le modèle. Le choix du modèle détermine quels types de données (D) rendre disponible pour le calibrer. Procéder enfin à la planification de la collectes des données (D) compatibles avec (M) et à l'exécution de l'expérience et appliquer le modèle (M) aux résultats de l'expérience, aux données générées par l'expérience.

Une attention particulière doit être portée au fait que **la collecte des données ne doit pas précéder le choix d'un modèle**. Et donc c'est le modèle qu'on ajuste aux données et non l'inverse.

Ceci Ici les restrictions commencent : Le modèle (M) ne sera valable que sous certaines conditions ou suppositions.

L'ANOVA n'est valable que pour les données issues des mesures prises dans les mêmes conditions. Dans le jargon des statisticiens on dit qu'elle est valable pour les données distribuées normalement avec la même variance.

L'ANOVA n'est donc pas appropriée pour les données d'observation en météo par exemple.

Donc, l'ensemble de données (D) ne sera utilisable que si ses éléments satisfont certaines propriétés.

La confrontation de (D) et (M) on note ($M \times D$) donne (M ajusté). Si (M) est bien ajusté aux données, alors passer à la phase III.

Sinon, modifier le modèle ou obtenir les données supplémentaires et recommencer. On obtient alors un nouveau modèle qui est confronté à la même question sur l'acceptation ou non. Si le modèle est accepté alors on passe à la phase III.

Quand est ce qu'on dit qu'un modèle est retenu ?

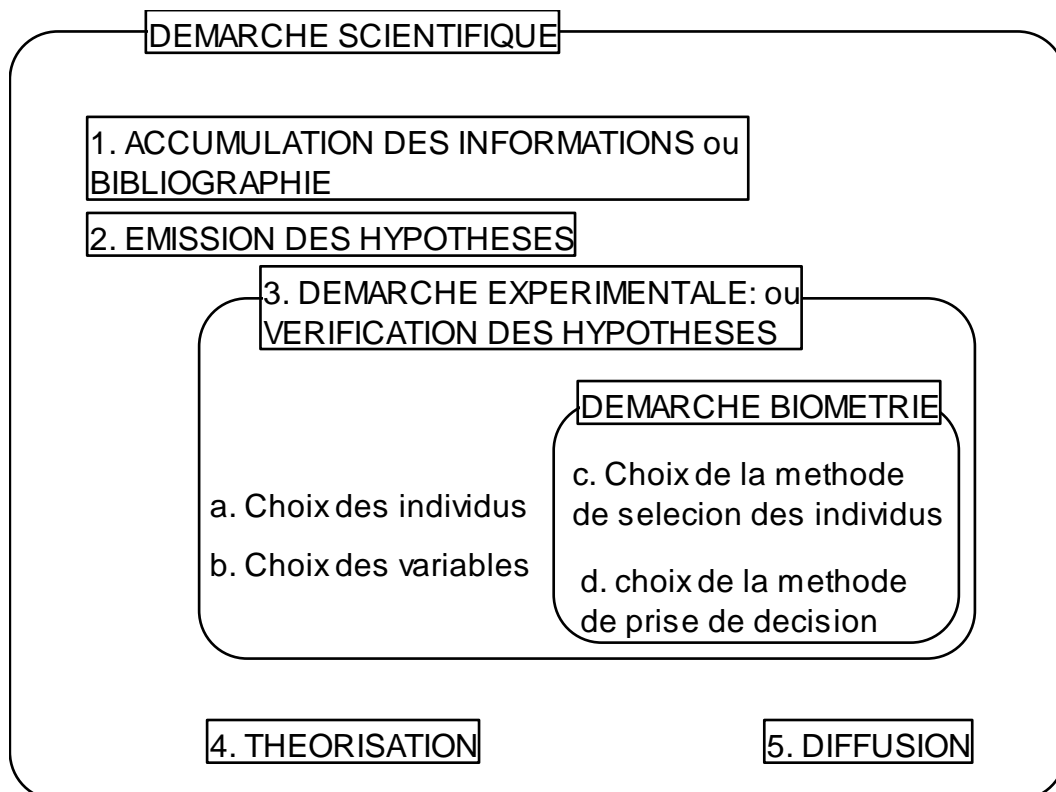
- Le modèle doit pouvoir reproduire les observations
- Il doit être simple avec un faible nombre des paramètres
- Il doit avoir une bonne adéquation aux objectifs de l'étude,
- Il doit permettre l'approfondissement des connaissances

Phase III. Simulation

Le modèle accepté est utilisé pour faire la simulation (S). A ce niveau, on confronte de nouveau le modèle à la réalité. Si le modèle est acceptable alors on continue à l'utiliser. Sinon on le modifie. Le résultat de cette modélisation est double : premièrement, on a la réponse à la question initialement posée et deuxièmement, on a l'accumulation de nouvelles connaissances.

Conclusion

Schématiquement, la démarche de la biométrie est contenue dans la démarche expérimentale qui elle-même est incluse dans la démarche scientifique.



C'est donc quand le chercheur décide de vérifier ce qui l'intrigue qu'il procède au choix des individus et des variables. Quand il commence à se poser la question comment choisir les individus, l'implication du biométricien commence. Quand il commence à prendre les mesures correspondant aux variables respectives et comment combiner ces

variables dans une formulation d'une problématique claire qui appelle une décision le besoin de la biométrie augmente.

Une synergie s'impose car le chercheur spécialiste en biologie ou en agronomie est souvent sans formation spéciale en biométrie et le biométricien est sans background suffisant dans le domaine mais peut proposer plusieurs pistes sur les techniques de prise de mesure ou de décision.

Les hypothèses biologiques sont vérifiées à travers des modèles statistiques complexes. Déterminer si un modèle choisi s'ajuste bien aux données collectées est un travail qui se fait correctement par un biométricien professionnel. L'apport du développement de l'informatique est la mise sur le marché des logiciels qui facilitent le traitement des données. Cette facilité s'accompagne de quelques exigences : L'utilisateur est supposé connaître les faiblesses et les conditions d'application d'un modèle particulier. Il doit être capable d'identifier le type de modèle approprié pour un dispositif particulier d'échantillonnage des données monté pour montrer un mécanisme dans un système biologique. Il doit être capable d'interpréter les sorties d'analyses en utilisant ces modèles. Il doit enfin savoir planifier une expérience de façon optimale.

Vu sous cet angle, les analyses doivent être faites par un biométricien professionnel et non par un biologiste formé en biométrie. Si ce biométricien est disponible, faut-il encore quelques conditions soient remplies : les questions doivent être formulées par le biologiste de sorte qu'elle soit perçue de la même façon par le biométricien.

Les exigences professionnelles du biologiste et du biométricien doivent converger. Cela n'est pas automatique. Par exemple, un défenseur des cultures a besoin que dans une expérience, il observe moins des feuilles attaquées par des insectes. Le biométricien quant à lui, a besoin, pour la même préoccupation, de compter plusieurs feuilles, au minimum trente afin que les conditions d'utilisation de ses méthodes soit remplies.

Le biométricien doit savoir retransmettre ses résultats en terme compréhensible par le biologiste.

