PROBLÉMATIQUES EXPÉRIMENTALES ET MÉTHODES STATISTIQUES

Y. ESCOUFIER

I.1) Il n'y a pas de **statistiques** sans données.

Les méthodes statistiques n'inventent pas de l'information. Elles ne peuvent que remplacer les données disponibles par des **résumés** qui ont l'avantage d'être **synthétiques** donc manipulables et le désavantage d'être **simplificateurs** : une série d'observations ne se réduit pas à sa moyenne ; un plan factoriel ne donne qu'une vision approchée d'une réalité plus complexe.

I.2) "Faire des statistiques" c'est présenter des résumés, des résultats et les commenter.

"Faire de la statistique" c'est développer des méthodes qui fourniront de nouveaux résumés et expliciter les avantages et les propriétés de ces résumés.

"Faire des statistiques" c'est par exemple calculer une moyenne arithmétique. Donner un intervalle de confiance pour cette moyenne, c'est à un certain sens la commenter.

"Faire de la statistique" c'est par exemple avoir trouvé dans un certain contexte la loi dite de Student qui permet la construction de l'intervalle de confiance dans ce contexte ; c'est étudier dans quelle mesure la construction reste valide si on s'écarte du contexte initial.

Les moyens informatiques rendent aujourd'hui tout à fait facile l'obtention des résultats. L'interprétation suppose une connaissance raisonnable des bases mathématiques des méthodes.

II.1) Pourquoi recueillir des données ?

- Pour prendre connaissance d'un phénomène, dans un but d'**exploration.**
On se demandera quelles sont les espèces d'arbres présentent dans une forêt ? On se demandera s'ils existent des liens entre la taille des arbres et leur position ?
On est là en présence de **Questions ouvertes.**
La liste des réponses possibles n'est pas fixée à l'avance.

- Pour vérifier des hypothèses faites sur le fonctionnement d'un phénomène. On veut **mettre une hypothèse à l'épreuve,** pour la confirmer ou l'infirmer.
On veut vérifier qu'il y a plus de pins que de bouleaux. On veut vérifier un effet nuisible des positions de bordure.
On est là en présence de **questions fermées**. La liste des réponses possibles est fixée à l'avance. Dans nos exemples, la réponse doit être Oui ou Non.

Il est important quand on entreprend une étude statistique de savoir dans quel contexte on se situe car les méthodes susceptibles de répondre à des questions ouvertes ne sont pas les mêmes que celles qui peuvent répondre à des questions fermées.
La problématique générale de l'étude doit guider dans le choix des méthodes à employer. La maîtrise d'une méthode statistique n'est pas une raison suffisante pour son emploi ; on doit se demander si elle est adaptée à la problèmatique de l'étude.

II.2) Qui fixe la liste des variables à observer? Qui dresse le questionnaire?

- Mise à l'épreuve d'une hypothèse : l'énoncé de l'hypothèse fournit le plus souvent la liste des variables concernées.
  Dans le premier exemple évoqué plus haut pour ce contexte c'est naturellement le nombre d'arbres qui doit être observé. Dans le second, on s'intéressera à des variables mesurant l'état des arbres : production, morphologie...

- Exploration : le désir d'exhaustivité conduit souvent à penser que la liste des variables doit être la plus longue possible. Il faut être conscient du fait que les inévitables connaissances préalables induisent des biais : il est toujours facile d'inventer des descripteurs pour un phénomène connu ; on peut totalement oublier un des aspects du phénomène.
  On doit recommander une réflexion a priori sur les différents aspects du phénomène au besoin en regroupant des experts d'origine différente.
  On demandera aux experts de dresser des listes des variables potentielles pour les différents aspects du phénomène.
  Une première étude statistique aura pour but l'étude critique de chacune des listes, en particulier la mise en évidence des redondances.

II.3) Statuts des variables observées

De même qu'on doit s'interroger sur la nature exploratoire ou confirmatoire de l'étude entreprise, on doit réfléchir aux rôles joués par les différentes informations disponibles.

- Certaines sont vraiment des descriptions des phénomènes étudiés. C'est leur comportement que l'on veut connaître.

- D'autres décrivent non pas les résultats obtenus mais l'environnement dans lequel le phénomène se déroule : ce sont des variables concomitantes dont on pense a priori qu'elles peuvent influer sur les résultats. Leurs valeurs seront parfois contrôlées par l'observateur, d'autres fois simplement observées.

Par exemple dans une étude portant sur la faune d'une rivière, les descripteurs du phénomène seront des nombres et des tailles de poissons.

Des variables concomitantes non contrôlées peuvent être des vitesses de courant, la température de l'eau, la profondeur de la rivière, la flore environnante. L'heure de la pêche, la nature du matériel de pêche pourront intervenir comme variable concomitante contrôlée.

Les études exploratoires oublient trop souvent cette réflexion sur le statut des variables attendant de l'ordinateur qu'il suggère des idées à partir d'un agglomérat de variables dont les statuts n'ont pas été précisés. C'est une mauvaise pratique. Toute méthode statistique cherche à fournir un résumé ; le résumé est d'autant plus pertinent et lisible qu'il résume des variables naturellement associables.

Des méthodes exploratoires existent qui permettent d'utiliser de façon explicite et active les informations concomitantes. Elles ont l'avantage de permettre une vision plus fine de la part du phénomène étudié qui n'est pas dû à leur comportement.

III) Méthodes confirmatoires

Principe :
Tout ensemble fini d'observations est issu d'une population hypothétique qui l'englobe.

L'objectif de toute étude est alors de façon explicite ou non d'induire du connu observé des informations sur la population hypothétique. Ceci fait, on pourra anticiper les comportements possibles d'autres finis éventuellement observables.

La méthodologie repose sur une caractérisation mathématique de la population. Dans l'approche la plus classique, la caractérisation comporte une forme mathématique dépendant de différents paramètres (la loi normale dépend de deux paramètres).

Le fini observé permettra d'**estimer** les valeurs des paramètres ou bien de les **comparer** à des valeurs théoriques données ou à d'autres valeurs expérimentales.

Celui qui fait de la statistique trouve dans ce contexte des champ d'études théoriques immenses : Il étudie les propriétés des estimateurs, leur comportement quand la taille de l'échantillon augmente, il définit de nouveaux tests...

Celui qui fait des statistiques doit s'interroger sur la vraisemblance de la caractérisation de la population hypothétique : la méthode a été développée en supposant que dans la population hypothétique la variable avait une distribution normale. Est-ce raisonnable dans cette application précise ?

Il doit s'interroger sur l'adéquation entre le fini qu'il a observé et le fini considéré dans les déductions mathématiques : nombre d'individus suffisamment grand ; individus indépendants...

De nombreux travaux récents (robustesse, approche non-paramétrique, méthodes de ré-échantillonnage) tendent par des chemins différents à détendre le cadre souvent très contraignant des méthodes classiques.

IV) Méthodes Exploratoires

Objectif : mettre en évidence :

- les ressemblances et oppositions entre individus,
- les liaisons entre variables;

La méthodologie consiste à remplacer les tableaux observés par des (tableaux voisins susceptibles de) représentations graphiques (exactes) la lecture des graphiques apporte une information (exacte sur les tableaux voisins qu'on utilise comme une information approchée) sur les tableaux observés.

Trop de praticiens lisent la phrase précédente en oubliant les parties mises entre parenthèses. C'est oublier que toute méthode simplifie pour résumer. Il sera toujours fondamental dans une méthode exploratoire d'apprécier l'approximation faite, c'est-à-dire la distance entre les données réelles et le résultat fourni.

Celui qui fait de la statistique rencontre à développer de telles méthodes des problèmes de nature souvent géométriques donc mathématiques mais aussi informatiques.

L'utilisateur doit être capable d'apprécier la qualité de l'approximation faite : c'est facile dans les méthodes factorielles qui apportent naturellement un élément quantificateur de cette approximation. C'est plus difficile pour les méthodes de classification et les méthodes récentes faisant un usage intensif de l'ordinateur.

Il doit aussi être capable de lire les résultats fournis c'est-à-dire qu'il doit avoir une bonne connaissance des règles de lecture des représentations graphiques.