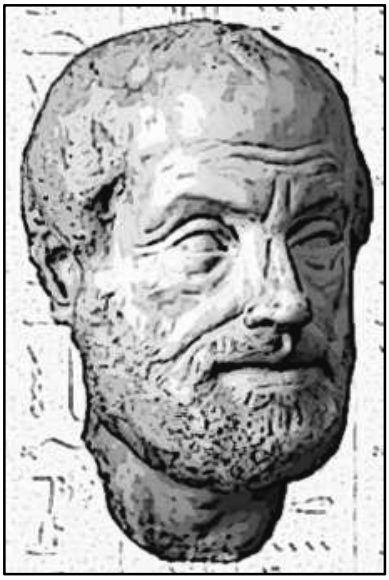
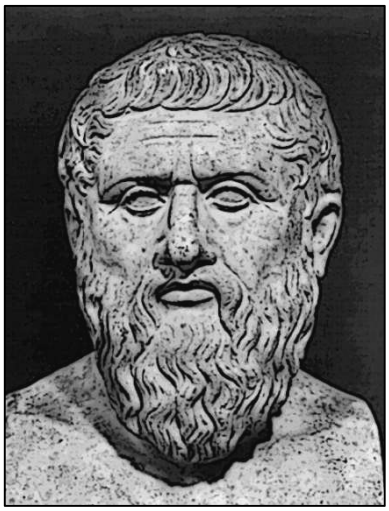
PARTIE 1
PREALABLES A UN TRAITEMENT STATISTIQUE



Aristote (-384 : -322)



Platon (-428 : -348)

John Graunt (1620 : 1674)

René Descartes (1596 : 1650)

CHAPITRE 1

UNE DEMARCHE SCIENTIFIQUE

La Statistique
« Voici les chiffres communiqués par les services de la statistique et intéressant la période comprise entre le 2 juillet et le 4 septembre : 543 285 ; 6 282 826 ; 1 285 938 743 ; 601 ; 602 ; 603 ; 604 ; 605 ; 106 ; 206 ; 306 ; 406 ; 506 ; 983 ; 882 ; 780 ; 680 ; 579.
Nous ne savons pas du tout à quoi se rapportent ces chiffres, mais nous sommes heureux de les communiquer à nos lecteurs qui auront ainsi toute latitude de les adapter suivant leur goût ou leur appréciation... »
Pierre DAC, *L'os à Moelle, Juillard, Paris (1963)*

Dans ce chapitre, nous allons présenter quelques idées qui vont guider notre approche de l'analyse statistique et de l'exploration des données. Après quelques réflexions sur les fondements de l'analyse statistique (§1), nous montrerons qu'en pratique une analyse n'est pas uniquement la réponse à une question mais qu'elle s'insère dans une démarche scientifique où il existe un aller-retour permanent entre la réflexion sur un problème et la réalité à laquelle nous sommes confrontés au sujet de ce problème : nous partons d'idées élémentaires que nous affinons progressivement (§2). Nous essaierons de formaliser cette démarche en montrant que cette formalisation nous oblige progressivement à nous poser des questions (§3). La possibilité de réaliser des analyses statistiques a été étendue à de nouveaux domaines dans lesquels la masse des informations à traiter était un écueil ; elle est désormais facilitée par l'existence de logiciels de calcul scientifique qui peuvent être facilement reliés à des logiciels de traitement de bases de données (§4).

1. PREALABLE A UNE ANALYSE STATISTIQUE

Est-ce que ce que nous voyons nous aide à mieux connaître le monde qui nous entoure ? En partie seulement, car notre vision si précise soit-elle n'est pas la **Réalité**, si tant est qu'elle existe, mais une version floue, bruitée. Pour **connaître**, il faut se poser au moins une question ; car la connaissance ne peut s'acquérir qu'en répondant à une question. La science statistique est constituée par des outils qui nous aident à mieux connaître le monde qui nous entoure en posant des questions puis en construisant des modèles traduisant celles-ci et dans lesquels nous introduisons une partie aléatoire. Comment parvenir à connaître le monde dans lequel les questions que nous posons peuvent avoir au moins une solution ? Deux approches sont possibles : doit-on partir d'une idée ou d'un fait ? Suivant ce que proposent Nisbet et *al.* (2009), on peut penser comme Platon ou comme Aristote : le premier croit que les choses les plus importantes dans notre existence sont au-delà de ce que nos yeux peuvent voir et que nos mains peuvent toucher ; le second croit que la connaissance d'un système complexe provient de

l'étude détaillée du milieu qui l'entoure et, pour en comprendre la réalité, nous devons décomposer ce système en morceaux que nous devons décrire séparément avant de les regrouper pour comprendre le fonctionnement global. Pour Aristote, la connaissance de l'ensemble est celle de la somme de ses parties, alors que pour Platon elle est supérieure à cette somme. La vision du premier n'est que très partiellement satisfaisante car dans un système global (comme un écosystème ou une entreprise commerciale) l'association des parties peut créer des propriétés nouvelles qui n'apparaissent que par l'existence de cette association. Techniquement ceci signifie que l'action de deux variables n'est pas toujours prévisible par la seule connaissance de chacune d'elle. C'est ce que Nisbet traduit en opposant *product-oriented knowledge-management* à *process-oriented knowledge management*. Nous en verrons des exemples tout au long de cet ouvrage.

Existe-t-il des domaines professionnels ou scientifiques desquels l'analyse statistique est absente ? Nous l'espérons. Néanmoins, ils sont plus difficiles à trouver qu'on ne le pense. Regardons autour de nous des domaines variés que rien ne semble rapprocher, sans aucun lien apparent, tous font appel à la statistique : de la physique théorique (où l'on passe du comportement microscopique de particules individuelles à un comportement macroscopique) à la datation de céramiques en archéologie, en passant par les Sciences Humaines et les Sciences de la Vie (linguistique, médecine, agronomie, écologie, etc.) et les Sciences Economiques et Sociales (l'économétrie bien sûr mais aussi la sociologie). Si nous nous aventurons vers d'autres domaines liés à l'Argent, l'intrusion est encore plus flagrante : jeux de hasard mais aussi assurance et finance et naturellement le marketing où le sondage d'opinion devient un outil pour prendre des décisions et faire des investissements. Dès lors que des informations chiffrées apparaissent et qu'il faut arriver à les maîtriser, la statistique appliquée fournit les outils techniques indispensables. Mais ne fournit-elle que des outils ? Comme d'autres l'ont dit avant nous, nous ne le pensons pas (Rao, 1994).

Analyser des données grâce à des outils statistiques est le but essentiel de la statistique appliquée. Mais pour analyser des données encore faut-il les relier à un problème que nous souhaitons étudier et à un ensemble de questions que nous nous posons au sujet de ce problème. Il faut aussi que ces questions ne soient ni vagues ni trop générales. Le statisticien travaille rarement en solitaire, il est la plupart du temps associé à une autre personne qui lui soumet un problème. Le statisticien doit donc aider celle-ci à préciser ses questions de manière à ce qu'il puisse y répondre. Pour y parvenir nous verrons qu'il faut suivre un chemin rigoureux. Ce chemin doit permettre de fournir une réponse, peut-être même plusieurs. Cette réponse est ce que l'on pourrait appeler la destination de l'analyse. Or, si arriver à destination est essentiel, bien souvent c'est le chemin suivi qui nous apporte le plus d'informations importantes en améliorant notre connaissance. C'est dans ce sens que la statistique appliquée n'est pas qu'un outil mais qu'elle peut devenir la base d'une démarche scientifique, donc rigoureuse par nature, du traitement de données avec un hasard toujours présent qui est une composante importante qu'il faut savoir apprivoiser.

Toute analyse est confrontée à trois obstacles :

1. Dispose-t-on d'un corpus de données suffisamment important ?
2. Ce corpus représente-t-il correctement l'univers à explorer ?
3. Ce corpus est-il pertinent pour répondre au problème posé ?

Si on a su franchir ces obstacles, il est alors possible de définir une stratégie pour résoudre ce problème. Les auteurs anglo-saxons parlent souvent de *process of data mining*. Nous allons étudier de façon plus détaillée cette stratégie : ce n'est pas une approche « linéaire » du problème, du type une question suivie d'une réponse. Partant de la question initiale, c'est plutôt un chemin le long duquel on s'arrête pour faire des vérifications, des *checkpoints*, éventuellement pour formuler différemment la question initiale. On avance donc de manière itérative vers une solution, si possible la meilleure, à la question posée.

2. SCHEMA D'UNE DEMARCHE SCIENTIFIQUE

Dans toute démarche scientifique, nous pouvons distinguer trois phases :

Phase 1 : pour atteindre le **but** $\mathcal{B}$ d'une étude, il est indispensable de partir d'un **bilan des connaissances** $\mathcal{C}$ déjà acquises au moins partiellement. Le but $\mathcal{B}$ peut être l'étude de la croissance de bactéries dans un produit alimentaire. Un **objectif** scientifique $\mathcal{O}$ peut alors être dégagé ; s'il doit être précis il n'a pas encore à se référer à une méthode particulière. L'objectif $\mathcal{O}$ peut être de prévoir le nombre de bactéries au bout d'un certain temps. Par contre, la définition de $\mathcal{O}$ implique que nous connaissions les grandes classes de questions auxquelles nous sommes capables de répondre. Une liste non limitative de questions possibles est : décrire des observations, expliquer des relations, estimer des paramètres, comparer deux ou plusieurs stratégies, classer des observations dans des groupes connus, classifier des observations, prédire le futur à partir d'observations passées, etc.

Phase 2 : l'objectif étant fixé, il est possible de faire l'inventaire des modèles connus susceptibles de l'atteindre. Par **modèle**, nous ne faisons pas obligatoirement référence à une présentation formelle et abstraite (Legay, 1997). Dire qu'un échantillon est représenté par sa moyenne peut déjà être considéré comme un modèle, simpliste certes, mais un modèle. L'un de ces modèles $\mathcal{M}$ devient le premier candidat. Son choix correspond bien souvent à notre propre expérience et à notre compétence ; mais aussi à nos capacités à le traiter et à l'appliquer. Ces dernières supposent que nous puissions faire les calculs qu'il demande et que des données $\mathcal{D}$ soient disponibles pour le calibrer. Si elles ne le sont pas, il faudra les obtenir en prenant des mesures, en faisant des enquêtes, des expérimentations, etc. Mais bâtir un modèle $\mathcal{M}$ et obtenir des données $\mathcal{D}$ va nous enfermer dans un cadre relativement rigide et restreint :

- $\mathcal{M}$ ne sera valable que sous certaines conditions que nous appellerons des **suppositions**[1] ;

[1] En anglais : *assumptions*. Notons que le terme d'*hypothèse* est souvent employé à la place de supposition ; dans un contexte statistique il existe une confusion possible avec la notion de test d'hypothèses, en anglais *hypothesis*.

- L'ensemble des données $\mathcal{D}$ ne sera utilisable que si elles satisfont à certaines propriétés.

Les méthodes statistiques que nous présenterons permettent d'étudier les propriétés intrinsèques de $\mathcal{M}$; confrontées aux données $\mathcal{D}$ elles nous permettent d'obtenir un nouveau modèle noté $\widehat{\mathcal{M}}$ qui s'appelle le **modèle estimé**. Par exemple, pour estimer le taux de survie de bactéries dans un aliment au cours du temps nous pouvons choisir pour $\mathcal{M}$ une droite ; les données $\mathcal{D}$ nous permettent d'obtenir l'équation précise, c'est-à-dire les valeurs numériques qui nous permettent de tracer la droite sur un graphique, c'est $\widehat{\mathcal{M}}$. Mais l'obtention de $\widehat{\mathcal{M}}$ n'a été faite qu'au prix d'un certain nombre de simplifications, par exemple le choix d'une droite *a priori*. Sont-elles acceptables ? C'est la dernière étape dite de *validation* $\mathcal{V}$ de cette phase ; elle aboutit à l'acceptation ou au refus de $\mathcal{M}$ et de sa réalisation $\widehat{\mathcal{M}}$. La démarche passe par une interrogation :

- Oui $\mathcal{M}$ est acceptable, nous pouvons continuer ;
- Non $\mathcal{M}$ n'est pas acceptable, il faut donc rebrousser chemin et modifier le modèle, peut-être aussi obtenir des données supplémentaires.

Nous remplaçons $\mathcal{M}$ par un nouveau modèle $\mathcal{M}^{(1)}$, voire de nouvelles données $\mathcal{D}^{(1)}$, et nous parcourons les mêmes étapes que pour le premier modèle. Au bout d'un certain temps, par améliorations successives, par itération, nous aboutissons à un modèle que nous finissons par accepter ; pour éviter une trop grande abondance de notations nous l'appellerons $\mathcal{M}$. On peut alors passer à la dernière phase.

Phase 3 : puisque le dernier modèle a reçu un quitus, nous allons l'utiliser comme *substitut* à la réalité. Il peut être utilisé pour prédire des situations nouvelles qui peuvent se présenter. Il peut aussi servir à simuler des situations non encore explorées. Ici encore, une confrontation avec la réalité est indispensable, généralement en vérifiant si ce qui a été prédit est suffisamment proche de la réalité observée. Là encore, si nous sommes satisfaits des résultats nous continuons à utiliser $\mathcal{M}$, sinon nous le modifions à nouveau.

Donc en fin d'étude, nous avons une réponse $\mathcal{R}$ à la question initialement posée et surtout l'accumulation de nouvelles connaissances : $\mathcal{C}$ est remplacée par $\oplus\ \mathcal{C}$.

3. D'UN BUT A UNE REALISATION

3.1 Triplet fondamental

Du schéma précédent, il ressort :

- Un *schéma itératif* de l'analyse du problème qui était posé. Nous avons développé un *processus d'apprentissage* ;
- Un triplet $\{\mathcal{O}, \mathcal{M}, \mathcal{D}\}$ qui constitue l'outil méthodologique permettant de répondre le plus correctement possible à l'objectif simplifié $\mathcal{O}$, réponse partielle au but $\mathcal{B}$.

Dans cet ouvrage, nous nous limiterons essentiellement à la seconde phase ; nous essaierons de montrer comment à un objectif $\mathcal{O}$ nous pouvons associer un modèle $\mathcal{M}$

choisi parmi d'autres. Un modèle est la vision à un instant donné d'une partie du monde réel. Ce n'est qu'une vision partielle, idéale dont on peut simplement dire qu'elle est utile pour atteindre l'objectif fixé. Chaque modèle implique l'existence de données $\mathcal{D}$; elles mêmes peuvent servir à des modèles différents. Il existe des relations entre les trois éléments de ce triplet (Fig.1). Bien sûr, tout ce schéma doit être complété par l'introduction de moyens informatiques qui rendent le traitement des données quasiment transparent pour l'utilisateur.

A un objectif $\mathcal{O}$, nous essaierons de faire correspondre au moins un modèle $\mathcal{M}$. Mais il peut en exister plusieurs ; ce sera d'abord au statisticien de choisir le plus adapté en fonction de ses compétences et des moyens dont il dispose. Il devra préciser les conditions de son utilisation, les limites de son exploitation par un utilisateur non-statisticien. Ce dernier doit intervenir car, beaucoup mieux que le statisticien, il connaît son domaine ; il connaît aussi les limites de compétence de ceux qui peuvent l'utiliser ultérieurement de manière régulière. Il faudra aussi comprendre qu'un bon choix des données $\mathcal{D}$ est essentiel : il est inutile de les accumuler en croyant qu'un ordinateur va faire découvrir par un coup de baguette magique des richesses insoupçonnées. Il sera donc utile de réfléchir à la meilleure façon de les utiliser, par exemple pas toutes en même temps. Mais il ne faut pas perdre de vue que la qualité des données $\mathcal{D}$ est un point extrêmement important. Si les données sont trop entachées d'erreurs, contiennent des effets liés aux traitements par lots[2], des biais dus à une mauvaise planification de leur recueil, le statisticien ne pourra pas faire de miracle ! La réponse $\mathcal{R}$ à la question posée sera de piètre qualité car aucun modèle ne parviendra à s'ajuster de manière satisfaisante aux données !

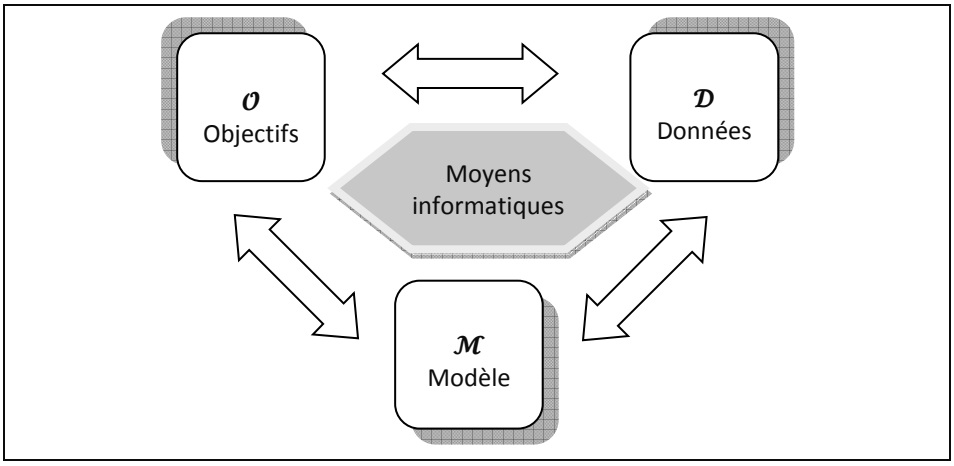


FIGURE 1. Etapes fondamentales d'une analyse statistique.

[2] Ce qu'on appelle des effets *batch* dans le traitement et l'utilisation de données de puces à ADN.

3.2 D'une vision idéale à la réalité

Nous allons essayer de préciser certains éléments de la démarche que nous avons décrite sur l'exemple de croissance de bactéries, sujet important pour garantir la sécurité alimentaire des produits que nous consommons. Si $\mathcal{M}$ est vrai, nous savons (c'est notre bilan de connaissances $\mathcal{C}$ qui nous permet de le dire) que la relation est simple, de forme quasiment linéaire $Y = a + bX$, où X est le temps (par exemple exprimé en jours) et Y le logarithme du nombre de bactéries. Si $\mathcal{M}$ était strictement linéaire, tout point (i) de coordonnées $Y = y_i$ et $X = x_i$ serait situé sur la droite d'équation $f(X) = a + bX$; il suffirait alors de deux points particuliers (deux observations) pour déterminer les valeurs de a et b. Mais, ce n'est sûrement pas le cas ; si on fait une expérience pour toute une série de valeurs (à des temps fixés par l'expérimentateur) $X = \{x_1, \cdots, x_n\}$ auxquelles on mesure le nombre de bactéries et que leur logarithme soit $Y = \{y_1, \cdots, y_n\}$, les n points de coordonnées $\{x_i, y_i\}$ forment un nuage plus ou moins proche d'une droite si $\mathcal{M}$ est raisonnablement correct. En règle générale, le modèle $\mathcal{M}$ n'est pas représenté par une relation mathématique, mais par une relation plus floue. Cette relation s'obtient en introduisant dans $\mathcal{M}$ une composante aléatoire ou bruit ou *aléa*. Plus précisément, ce que l'on observe est que Y (la variable à expliquer appelée quelquefois la réponse) est lié fonctionnellement à X (la variable explicative que nous appellerons plus loin régresseur[3]) par une relation plus complexe que $f(X)$ de la forme :

$$\mathcal{M}: "Y = f(X, paramètres, aléa)"$$

Sous cette forme le modèle n'est pas « opérationnel », il n'est pas encore utilisable pour faire une prévision ; c'est un « instrument de discours ». Il fait apparaître les éléments importants de la *phase 2* décrite plus haut :

- Y: la valeur observée (le logarithme du nombre de bactéries), image déformée d'une valeur hypothétique ;
- la fonction f qui relie Y à :
  - une quantité X, valeur d'une variable qui définit la **condition expérimentale** pour laquelle la valeur a été observée ;
  - des **paramètres** pour l'instant inconnus intervenant dans la fonction ;
  - l'**aléa**.

Nous noterons l'ensemble des p paramètres (ici $p = 2$) sous la forme d'un vecteur $\boldsymbol{\Theta} = (\theta_1, \cdots, \theta_p)^{\mathrm{T}} = (a, b)^{\mathrm{T}}$ et les n aléas (un pour chaque observation) par $\mathbf{E} = (\varepsilon_1, \cdots, \varepsilon_n)$. Nous pouvons donc réécrire formellement le modèle :

$$\mathcal{M}: "Y = f(X, \boldsymbol{\Theta}, \mathbf{E})"$$

Jusqu'ici rien n'est encore opérationnel ; pour avancer nous pouvons nous demander comment intervient l'aléa ; il est souvent commode de supposer que son effet est **additif**, et qu'alors $\mathcal{M}$ peut s'écrire :

$$\mathcal{M}^+: "Y = f(X, \boldsymbol{\Theta}) + \mathbf{E}"$$

[3] Plus généralement il peut y avoir plusieurs régresseurs.