STATISTIQUE ET HASARD (passim, O)
(23 / 03 / 2020, © Monfort, Dicostat2005, 2005-2020)

*« Une **intelligence** qui, à un instant donné, connaîtrait **toutes les forces** dont la nature est animée, **la position respective des êtres** qui la composent, si d'ailleurs elle était assez vaste pour soumettre ces données à l'analyse, embrasserait dans la **même formule** les **mouvements** des plus grands corps de l'Univers, et ceux du plus léger atome. **Rien ne serait incertain** pour elle, et **l'avenir comme le passé seraient présents à ses yeux** »*
Pierre Simon, marquis de Laplace (1749-1827)

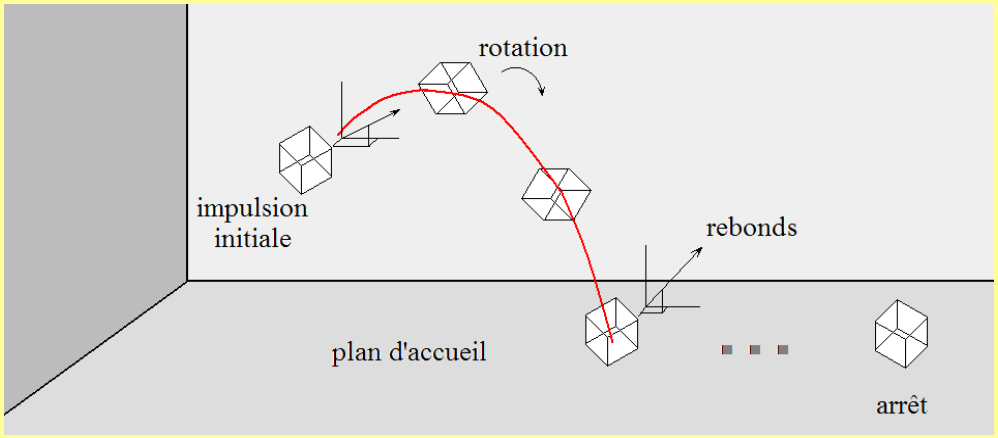
(i) Probabilité et hasard

Le **calcul des probabilités** constitue un moyen pour approcher la compréhension d'un **phénomène** complexe, et aussi pour en « raccourcir » l'analyse scientifique (cf **étude scientifique**).

Ainsi, dans la « simple » expérience du **lancer de dé** (à six faces) **sur un plan**, le **résultat** d'un lancer donné (qui peut comporter des conséquences d'ampleur : eg quitte ou double) est a priori inconnu, même s'il ne peut prendre ses valeurs que dans la suite des nombres entiers $N_6^* = \{1 ,..., 6\}$.

(a) on peut tenter de « prévoir » ce résultat en utilisant les « **lois** » qui sont supposées gouverner le phénomène précédent (physique) : intensité x_1 et direction x_2 de la force initiale (impulsion), vortex x_3 imprimé par la main sur le dé, frottements x_4 dans l'air ou d'autres obstacles éventuels, rebondissements sur le plan (qui dépend des élasticités x_5 et x_6 des matériaux constituant le dé et le plan), etc.

Le résultat y résultant du tirage dépend donc, selon une relation complexe notée f, des variables tq (x_1 ,..., x_6). Chacune de ces variables doit, en outre, être considérée comme aléatoire : eg x_1 serait uniformément distribuée sur un segment [a, b], x_2 suivrait une loi directionnelle, x_3 dépendrait de x_1, etc (cf schéma ci-après).



Cette approche rencontre cependant deux limites :

$(a)_1$ l'**état d'avancement des sciences** (point de vue théorique). En effet, la formulation des **lois scientifiques** (en l'espèce, celles de la physique) peut toujours évoluer et s'améliorer. Le résultat d'un calcul « théorique » (ie ex ante, avant tout lancer) dépendrait

donc d'abord de ces lois (même si, dans le cas élémentaire du dé, ces lois sont généralement bien connues) ;

$(a)_2$ les **imperfections altérant l'expérience** (point de vue pratique). Elles sont dûes à la difficulté pratique de prendre en compte tous les facteurs qui influencent le phénomène du lancer de dé. En particulier, divers défauts sont inhérents au milieu ambiant (air et ventilation, sol et régularité) ou aux objets utilisés (contrôle de la main, dé). Le matériau constituant le dé n'est pas « homogène », les nombres représentés sur sa surface (peints ou creusés dans la matière) sont différents : donc la répartition de la masse du dé n'est pas symétrique par rapport à son centre de symétrie supposé. De même, l'arrondissage des angles et des sommets est imparfait. Le plan d'atterrissage peut aussi souffrir de défauts de planéité (surface gauche) ou d'homogénéité. Enfin, l'expérimentateur lui-même n'est pas exempt de défauts, ce qui justifie de considérer comme aléatoires les différentes variables pouvant intervenir dans l'expérience ;

(b) il paraît alors plus simple et plus rapide d'utiliser les propriétés du calcul des probabilités pour « prévoir » le résultat du lancer de dé, même si, dans l'exemple considéré, la probabilité d'une prévision juste peut paraître dérisoire (1/6) # 16,7 %.

En effet, dans toutes les sciences, on observe des situations concrètes beaucoup plus complexes (cf **complexité**) ; de plus, l'état d'avancement même de ces sciences, déjà inégal entre elles, rend pertinent le développement d'une approche stochastique.

Cependant, en passant de la démarche (a) à la démarche (b), l'objectif (ou l'ambition) recherché(e) a changé :

(a) la démarche (a) est de type complet et analytique, puisqu'elle étudie le phénomène depuis ses prémisses (lancer) jusqu'à son résultat (numéro sorti), en passant par la description et l'enchaînement des diverses « **lois scientifiques** » pouvant être mises en jeu pendant son déroulement. Or, ces lois ne peuvent résulter que d'estimations statistiques : elles sont donc, par construction, aléatoires ;

(b) la démarche (b) se contente d'analyser le « résultat » du phénomène (lancer du dé) in fine, en utilisant seulement les caractéristiques géométriques (théoriques) du dé. Rien n'empêche cependant d'utiliser les défauts indiqués pour affiner l'analyse probabiliste (ainsi, dans le lancer d'une pièce de monnaie, on peut modéliser le côté face F, le côté pile P et la « tranche » T de la pièce).

(ii) Hasard et nécessité

Ce qui précède montre que la distinction classique entre ***hasard*** (résultats dénués de déterminismes apparents) et ***nécessité*** (les « lois » supposées de la Nature) n'aurait (peut-être) pas lieu d'être si toutes les lois étaient entièrement connues et si tous les « objets » entrant dans le déroulement des phénomènes étaient « parfaits » : dans ce cas, tout serait nécessité, et aucune place ne serait laissée au hasard (dans ce sens, cf philosophie laplacienne en exergue).

Ce n'est donc que dans la mesure de la connaissance (plus ou moins avancée) des lois « gouvernant » les phénomènes observés que les « pronostics » qui leur sont relatifs seront plus ou moins exacts. Il existe donc une sorte de ***« balance » entre hasard et nécessité*** : plus le scientifique peut « réduire » le rôle du hasard dans la compréhension d'un phénomène, et plus les comportements phénoménaux observés relèveront d'une nécessité.

D'un point de vue formel, on considère :

(a) $\forall\ i \in I$, un **schéma probabiliste** $(\mathcal{X}^{(i)}, \mathcal{B}^{(i)}, P^{\xi(i)})$, comportant une **famille** $\xi(i)$ de variables décrivant un phénomène donné avec une **loi de probabilité** $P^{\xi(i)}$ (cf **loi multivariée**, **relation fonctionnelle**) ;

(b) ξ la famille $(\xi(i))_{i \in I}$ constituée de telles familles.

On suppose qu'il existe dans I un cheminement (sans boucle) de i vers un **indice** i_∞ tq les familles $\xi(i)$ correspondantes rendent les lois $P^{\xi(i)}$ de plus en plus « précises », jusqu'à approcher une **loi de DIRAC** $\delta_{(\xi(i\infty))}$ chargeant le « point » $\xi(i_\infty)$, où i_∞ (noté par commodité $i\infty$) désigne la « famille limite » des variables possédant cette propriété.

On peut alors considérer que la situation statistique a évolué depuis une situation « initiale » dotée d'une certaine incertitude vers une situation « finale » dans laquelle tout est déterminé.

Cependant, on peut aussi admettre qu'un phénomène puisse posséder une **variabilité** « intrinsèque » (éventuellement importante), qu'une meilleure compréhension ne pourrait cependant jamais réduire.

On peut rapprocher l'attitude précédente d'une approche statistique courante, dans laquelle on recherche une « **représentation statistique** » (ou « modèle statistique ») en **adéquation** aussi forte que possible avec des observations (les « données ») (cf aussi **loi multivariée**).

Dans l'exemple d'une **régression**, une liste (supposée finie) $\xi = (\xi_1, ..., \xi_K)$ de **variables exogènes** est censée influer sur une liste (aussi finie) de **variables endogènes** $\eta = (\eta_1, ..., \eta_G)$ selon l'équation usuelle (forme explicite à **perturbation aléatoire** additive, écrite dans l'**espace des variables**) :

(1) $\quad \eta = f(\xi) + \varepsilon,$

la perturbation ε exerçant une influence nulle en « moyenne » sur les variables de gauche.

Dans le **domaine de connaissance** concerné, le **statisticien** détermine les listes ξ, η et spécifie le type analytique de la fonction f. Il peut alternativement adopter une **méthode non paramétrique**. Un modèle plus « complexe » (ou « évolué ») que (1) serait :

(2) $(\eta, D\eta) = (f + Df)(\xi + D\xi) + (\varepsilon + D\varepsilon)$,

où les listes de variables sont symboliquement indiquées sous la forme $\zeta + D\zeta$ pour signifier que la liste $\zeta + D\zeta$ contient la liste ζ, et où la fonction $(f + Df)$ « généralise » (ou « étend ») la spécification de la fonction f.

(iii) Statistique et phénomènes aléatoires

Si le « vrai » modèle est (1), et qu'il traduit la « vraie » **causalité**, alors :

(a) la **nécessité** (le **signal**) est formalisée à travers f ainsi que la liste (ξ, η) des variables concernées ;

(b) le **hasard** (le **bruit**) intervient sous la forme « résiduelle » de la perturbation ε.

La perturbation ε est donc considérée comme une **altération** du « lien » f. Si l'influence de ε est faible, le signal est bien perçu et l'utilisation du modèle (1) (en prévision, notamment) ne pose pas de problème particulier. Si l'influence de ε est forte, le bruit est bien perçu et l'intérêt du modèle en est d'autant plus limité : le statisticien cherche souvent, dans ce cas, à repérer dans ε l'effet d'une possible omission de variables importantes (cf aussi la relation entre la normalité de ε et la **loi des grands nombres**).

Si le modèle (1) est imposé, l'altération précédente dépend alors de la « variabilité » de ε, laquelle est « intrinsèque » et ne peut, si les hypothèses retenues sont vraies, être modifiée par le statisticien : c'est le domaine de connaissance considéré, ainsi que la nature du phénomène observé, qui peuvent expliquer une plus ou moins forte variabilité.

Par contre, si le modèle (1) n'est pas imposé (eg recherche de spécification), il arrive souvent, en pratique, que le statisticien cherche à « réduire » l'influence de ε en modifiant les autres entités du modèle (ξ, η et f).

(iv) Du caractère partiel de l'analyse scientifique

L'analyse d'un phénomène implique souvent une prise en compte aussi exhaustive que possible de la liste des variables pouvant décrire ce phénomène, ou dont certaines influent même sur ce dernier. En effet, le statisticien ne dispose pas toujours de l'ensemble des variables souhaitables, donc des observations y afférentes (cf **phénomène**) :

(a) soit que ces variables / observations relèvent d'autres phénomènes constituant le **domaine de connaissance** considéré ;

(b) soit que ces variables / observations relèvent de phénomènes constituant d'autres domaines de connaissance.

Ceci peut résulter de divers facteurs : division du travail entre statisticiens appartenant à des « sphères » d'activité différentes, conception imparfaite du modèle associé au phénomène (eg non détection ou omission involontaires de variables), mais aussi impossibilité d'observation, ou encore augmentation de complexité de **modélisation** que cela implique.