DICTIONNAIRE
DE LA STATISTIQUE ET DE LA SCIENCE
2005

PRÉSENTATION

(12 / 03 / 2025, © Monfort, Dicostat2005, 2005-2025)

RETOUR A L'ACCUEIL **VERSION ANGLAISE**

La **Statistique** constitue par excellence un ***intrument scientifique puissant*** qui permet d'établir le nécessaire *aller - retour entre les théories et les faits*, c'est-à-dire la *navette entre les hypothèses et les observations*, ou encore la *confrontation entre les idées et les « données d'observation »*. Cette caractéristique majeure d'une ***science***, habituellement réputée austère et difficile, en fait ainsi une démarche générale pour l'étude de l'***Homme*** et des divers « milieux » par rapport auxquels il cherche à se situer.

Pour des raisons pédagogiques, et avec un risque d'arbitraire, on distingue dans ce travail cinq milieux au sujet desquels l'homme peut vouloir satisfaire sa curiosité (compréhension) ou sur lesquels il peut vouloir agir (action) : il s'agit des milieux **physique**, **biologique**, **écologique**, **psychologique** et **sociologique** (cf ***domaine de connaissance***).

De façon sommaire, on admet ici que (cf aussi ***Grand mystère de l'existence***) :

(a) le **milieu physique** (physique du globe et des corps célestes) constitue un « substrat » général ;

(b) sur ce substrat s'est « greffé » au cours du temps un **milieu biologique** (faune, flore) ;

(c) les intéractions, aussi bien internes qu'externes, entre les milieux physique et biologique ont conduit à un ensemble complexe de relations définissant le **milieu écologique** ;

(d) le monde vivant a aussi développé des comportements qui définissent l'objet d'un **milieu psychologique** ;

(e) les groupes ainsi constitués ont développé divers modes d'organisation définissant un **milieu sociologique**.

On situe donc l'Homme comme étant à l'origine de divers questionnements relatifs à ces domaines, donc portant sur les divers **phénomènes** observables relevant de ces domaines.

Cependant, pour des raisons historiques, et aussi en raison du caractère spécialisé de l'activité statistique, une « division du travail » s'est imposée : une sorte de

« ***contrat social*** » a conduit l'Homme (ou ses représentants) à donner au scientifique - donc, à un degré ou à un autre, au statisticien - un mandat en vue de l'éclairer :

(a) sur la **nature** et sur le **fonctionnement** de ces milieux ;

(b) sur leur **passé** ou leur **futur** (évolutions attendues) ;

(c) enfin, sur les possibilités d'**action** sur ces milieux.

Il est important de souligner que cette connaissance n'est pas seulement spéculative ou désintéressée. Elle doit surtout apporter des informations (ou des guides) permettant de mieux agir sur les milieux en question. Il s'agit, en effet, de mieux satisfaire divers **besoins humains** en rapport avec les cinq domaines précédents, afin de gérer :

(a) l'espace physique (utilisations, modifications, extensions) ;

(b) la santé (non seulement humaine, mais aussi animale et végétale) ;

(c) les ressources naturelles et l'environnement, compte tenu de l'activité anthropique ;

(d) les comportements, individuels ou de groupe ;

(e) les organisations (administrations, entreprises, associations).

1. ***Définir la Statistique*** revient donc à ***préciser sa place dans le champ de l'activité scientifique***.

On peut d'abord justifier son **utilité** par son implication dans les divers domaines de l'activité humaine évoqués ci-dessus (physique, biologie, écologie, psychologie ou sociologie). Chaque domaine de connaissance a, en fonction des besoins ressentis, engendré des sous-domaines spécialisés intégrant, de façon étroite, les apports de la Statistique. Ainsi, les développements les plus récents, au regard de l'histoire des sciences, concernent :

(a) la physique : physique statistique, météorologie, climatologie, géologie ;

(b) la biologie : biométrie (ou biostatistique) ;

(c) la écologie : synécologie, écosystémique ;

(d) la psychologie : psychométrie ;

(e) la sociologie : linguistique quantitative, sociométrie, économie (économétrie, recherche opérationnelle), industrie et ses technologies (technométrie), assurance (actuariat), démographie (démométrie), etc.

Inversement, ***chaque domaine de connaissance a contribué à la naissance ou au développement de concepts, méthodes ou résultats spécifiquement statistiques***. Ceux-ci étaient parfois, au départ, propres au domaine. Mais ***l'expérience a montré qu'ils pouvaient souvent être transposés ou étendus avec fruits à d'autres disciplines***, donc acquérir ainsi une sorte d'autonomie et devenir des notions à la fois abstraites et fécondes.

En effet, de façon générale, le ***progrès scientifique*** procède par consolidation des trois piliers suivants :

(a) l'***amélioration des appareils d'observation statistique****, ou* ***dispositifs d'observation****, spécifiques de chaque domaine* (instruments d'observation plus « fins ») (cf ***système d'observation***). Ainsi (physique), l'usage des satellites permet de mieux observer les phénomènes météorologiques ou climatiques, par rapport aux relevés au sol ou par ballons-sondes antérieurs. Les sondes spatiales permettent de mieux observer l'univers et d'en repousser, dans une certaine mesure, la profondeur du champ d'observation. De même (biologie), l'imagerie médicale (rayons X, puis scanner et RMN) a amélioré l'observation des pathologies, qui était auparavant principalement « clinique », ainsi que les procédés d'endoscopie et de traitements (qui possèdent progressivement un caractère moins invasif ou moins traumatisant). Ou encore (écologie), le baguage de la faune a permis de mieux en connaître la démographie ou les migrations ;

(b) une ***meilleure utilisation des méthodes existantes***. Ainsi, l'existence de points aberrants (cf ***aberration***) dans certaines statistiques fausse les analyses standards (eg calcul de moyennes, estimation de régressions, prévisions). Or l'interprétation correcte de ce problème (***mélange de lois***, **lois à queue épaisse**, **erreurs de spécification**) a permis de concevoir des tests de détection, utilisés simultanément avec des méthodes robustes (cf ***robustesse***), pour pallier ces inconvénients ;

(c) le ***développement de méthodes nouvelles***. Ces dernières sont souvent produites par **associations d'idées** (concepts ou méthodes). Ainsi, l'analyse des ***séries temporelles*** non stationnaires (cf ***processus non stationnaire***) a conduit à l'élaboration de notions (eg **donnée intégrable**, **intégration**, **cointégration**) aptes à traiter ces particularités, tout en utilisant des concepts et méthodes dérivés de l'analyse des séries stationnaires (cf **processus stationnaire**, **stationnarité**).

Ce sont donc les **trois piliers** précédents que la Statistique contribue à renforcer.

2. ***Par suite, il paraît très justifié que la connaissance statistique soit plus poussée et aussi mieux répartie*** dans les différentes sphères de l'activité humaine. Cette justification est aussi une nécessité.

2.0. En effet, des ***limites importantes existent***, encore aujourd'hui, en matière de connaissance et de compréhension statistiques, tant au niveau de la formation initiale (enseignement) que de sa complétion ultérieure (formation approfondie, formation professionnelle). Un effort, personnel aussi bien que public, visant à améliorer cette connaissance ne peut que dynamiser la recherche scientifique.

En bénéficiant des effets attendus d'une telle « révolution », chaque pays pourrait engendrer des progrès, économiques ou autres, aidant à résoudre diverses difficultés, contre laquelle les démarches usuelles semblent limitées. Il s'agit, en effet, de ***difficultés structurelles et récurrentes***, qui peuvent donc être de nature : physique (mouvements du sol ou du sous-sol : séismes, tsunamis), biologique (pandémies nouvelles), écologique (activité anthropique et dégradation environnementale massive), sociologique (surpopulation mondiale, disparité économique intra- et inter-nationale, lutte pour l'accès aux ressources naturelles, voire intégrisme, belligérance, etc).

Aujourd'hui, on est très fondé à penser que seule une accélération du progrès général peut sérieusement aider à contourner ces difficultés. Il est donc urgent de ne pas attendre, notamment en raison de l'existence de particularités susceptibles de limiter l'action publique :

(a) ***hystérésis*** (effets-retards) : ainsi, une formation statistique individuelle ou l'amélioration d'un dispositif d'observation demandent du temps ;

(b) ***entraînements*** (effets en chaîne). Ainsi, la surpopulation mondiale entraîne un excès d'activité anthropique induisant : une surconsommation de ressources naturelles, dont notamment le stock renouvelable qui ne se renouvelle pas assez vite (quantitativement aussi bien que qualitativement), ainsi qu'un accroissement trop massif d'externalités négatives (pollution : GES, radioactivité, poussières et autres toxiques) dans les divers milieux et selon des vitesses de propagation très différentes (sols et sous-sols, milieux aqueux, air), etc ;

(c) ***irréversibilités***, nombreuses dans ce type de situations. Ainsi, les pertes de biodiversité peuvent priver l'humanité de moyens (molécules naturelles) pouvant conduire à développer diverses solutions ou à résoudre diverses urgences (alimentation, santé).

On rencontre aussi de nombreuses limites au niveau des mentalités, des ***préjugés répandus dans le « public »***. Ainsi, on doit corriger des **idées reçues** fréquentes, telles que les suivantes.

2.1. La Statistique est censée ne pouvoir traiter que des ***observations nombreuses***.

Cette opinion ignore qu'il existe des possibilités pour valoriser des données « rares » :

(a) dans certains cas, la **loi** sous-jacente à l'observation d'un ***phénomène*** est connue ou peut se déduire d'autres lois connues. Un postulat statistique de base est, en effet, le suivant : la connaissance de cette loi équivaut à la complète connaissance du phénomène. Ceci est généralement la situation en **calcul des probabilités** ou en matière de **simulation** (cf aussi ***nombre au hasard***) : dans ce contexte, la loi est soit connue (donnée), soit définissable à partir de ses propriétés mathématiques. Cependant, la situation statistique usuelle veut que la loi gouvernant le phénomène est, en général, inconnue : ceci justifie l'intérêt et la nécessité de l'**inférence statistique** ;

(b) dans d'autres cas, une ***répétition***, ou un ***renouvellement***, du phénomène sont réalisables (**expérimentation**), ce qui permet d' « atteindre » les lois intervenant dans le mécanisme de ces phénomènes (**plans d'expérience**, **propriétés asymptotiques**, **loi des grands nombres**, **problème de la limite centrale**) ;

(d) par ailleurs, la vérification (**test**) d'hypothèses de ***stationnarité*** permet de limiter l'analyse à une seule observation (au lieu de plusieurs, en général) (cas de la ***trajectoire*** d'un ***processus***) (cf ***monographie***) ;

(d) enfin, la ***théorie bayésienne*** est parfois applicable : celle-ci pallie, d'une certaine manière, la « rareté » de l'***information*** à l'aide d'une « croyance a priori ». Cependant, cette dernière n'est pas « gratuite » : elle peut résulter d'autres travaux similaires (antérieurs ou contemporains), ou encore servir à vérifier la résistance de certaines hypothèses par rapport à des jeux d'**observations** fictives (cf **échantillon artificiel**).

2.2. La Statistique n'est fondée que sur des ***situations expérimentales***, par nature répétitives, et non par des situations de type « historique » (donc, par essence, « uniques »). Ne serait alors « scientifique » qu'une activité fondée sur l'expérimentation.

Il est vrai que l'expérimentation (cf ***plan d'expérience***) constitue un mode statistique d'obtention de données classique dans les sciences réputées « expérimentales » : physique, biologie, écologie (agronomie). Mais il existe d'autres moyens d'investigation (statistiques tq les **sondages**, ou non statistiques tq les relevés d'observation divers), assortis de méthodes d'analyse adaptées, pour répondre aux questions relevant de sciences dites « non expérimentales » (cf **production statistique**).

On observe que les procédés d'expérimentation interviennent de plus en plus dans des sciences non expérimentales, et inversement. Autrement dit, la confusion, encore faite (ou entretenue), entre science et expérimentation tend à s'estomper au profit d'une démarche scientifique plus générale et plus systématique, au sein de laquelle la Statistique a largement attribué son style.

Dans le même ordre d'idées, une croyance, associée à la précédente, relie ***scientificité et précision*** (voire exactitude) : une science ne mériterait son titre que dans la mesure où la précision de ses résultats ou de ses prévisions serait « importante ». Or, une connaissance, même sommaire, de la Statistique et de son contexte infirme une telle croyance : ainsi, les « lois » de la physique peuvent parfois être moins concisément cernées (cf **processus du mouvement brownien**) que certaines lois psychologiques (eg procédés de conditionnement) ou sociologiques (eg comportements de sélection adverse ou d'aléas moral en économie).

2.3. La Statistique permet d'analyser des ***informations quantitatives*** (données numériques) mais non des données de nature « qualitative ».

Cette croyance est démentie par une distinction, classique en Statistique, entre ***variable qualitative*** et ***variable quantitative*** (concepts de base du statisticien).

Diverses notions connexes (***codage***-décodage, ***modèle qualitatif*** ou ***modèle mixte***, etc) permettent d'ailleurs de passer d'un type de données à l'autre.

En outre, dans la mesure où l'une des finalités de la ***science*** (donc, nécessairement, de la ***Statistique***) est de répondre « in fine » aux questions que l'Homme peut se poser, il paraît nécessaire d'introduire et de conclure un travail scientifique, généralement formalisé et rédigé en « jargon professionnel », à l'aide d'attendus ou d'explications intelligibles par l'Homme lui-même, c'est-à-dire par une personne n'ayant aucune connaissance particulière de la matière et de ses méthodes. Autrement dit, malgré la difficulté pédagogique qui se présente, il semble hautement souhaitable de parvenir à tenir un discours purement littéraire (donc habituellement de nature qualitative) lorsqu'une étude scientifique est présentée au mandant initial (cf contrat social évoqué supra).

2.4. La Statistique est de la ***mathématique appliquée***.

En fait, il s'agit d'une ***science dont la situation est très particulière*** et dont il convient de clarifier le statut.

2.4.1. **D'une part, la Statistique constitue une branche des mathématiques** (ou de la Mathématique) **à part entière**, puisqu'elle utilise des raisonnements de type « hypothético-déductif » ainsi que les outils mathématiques fondamentaux (logique, théorie des ensembles, algèbre, analyse, etc). **Elle ne constitue donc pas un savoir « appliqué »**.

Par ailleurs, si le calcul des probabilités s'intéresse à une loi de probabilité connue (ou « théorique ») et en étudie les propriétés mathématiques (cf transformation des lois, processus, etc), la Statistique comporte un angle de vue différent, voire inverse (raisonnements de type « déduco-hypothétique »). En effet, comme cela a été indiqué, elle suppose, de façon très générale :

(a) que tout phénomène est gouverné par une loi de probabilité (la « vraie » loi, en principe, inconnue) ;

(b) que cette vraie loi appartient à une **famille** de lois possibles (ou lois « potentielles ») ;

(c) c'est l'observation du phénomène en question qui permet, ***« par remontée »***, de détecter et de préciser cette vraie loi au sein de cette famille (cf **inférence statistique**).

En sens inverse, le calcul des probabilité, aussi bien que la Statistique, enrichissent d'ailleurs les autres branches des mathématiques (algèbre, analyse, géométrie) : tous deux généralisent le « déterminisme » mathématique par plongement dans un contexte aléatoire. Ainsi, le **calcul des probabilités** peut transformer une équation différentielle en une ***équation différentielle stochastique***, une équation intégrale en une ***équation intégrale stochastique***, la géométrie en ***géométrie stochastique*** (cf aussi **probabilité géométrique**), etc. La Statistique peut alors s'inscrire à la suite de ces concepts.

En effet, si le calcul des probabilités étudie les propriétés stochastiques des objets précédents, la Statistique en étudie les propriétés inférentielles : l'**observation** de **phénomènes** appropriés informe sur la loi de probabilité de ces objets aléatoires.

On peut illustrer quelques « transitions » entre contexte « déterministe » (mathématique) et contexte « stochastique » (probabiliste) à l'aide d'***exemples élémentaires***.

Ainsi, en tant que branche des mathématiques, la Statistique ne possède pas de signification particulière de type « application » : cet aspect est souvent ignoré du public, qui confond généralement cette science avec les ***faits chiffrés*** qu'elle contribue à construire (***production statistique***) ou à analyser (***étude scientifique*** et ***diffusion de l'information***). On doit ainsi rappeler la distinction de base entre :

(a) **LA Statistique** (ie la science) ;

(b) **UNE *statistique*** (ie le concept technique) ;

(c) **LES *statistiques*** (ie informations ou données, qualitatives ou chiffrées).

La confusion provient aussi de l'assimilation implicitement faite par le public entre « la » Statistique et « les » statistiques relatives à des domaines concrets : données météorologiques, évaluations démographiques, salaires et prix, etc.

On doit, en outre, remarquer que les mathématiques, aussi bien que la Statistique, ont aussi, très généralement, eu pour origine l'**analyse de problèmes concrets**. Ainsi en est-il :

(a) de l'**algèbre** et des **dénombrements** :

$(a)_1$ démographie : comptage de populations ou de soldats ;

$(a)_2$ ressources d'autorité : fiscalité, monnaie ;

$(a)_3$ questions techniques : pesage et unités de mesures, changements d'unités ;

(b) de la **géométrie** et de l'étude de **figures** dans l'espace :

$(b)_3$ astronomie ;

$(b)_2$ formes et techniques architecturales ;

$(b)_3$ cartes et plans.

Ces deux savoirs, initialement provoqués par des besoins de résolution de problèmes concrets, se sont ensuite organisés (systématisation, formalisation, axiomatisation, etc) pour constituer des « outils » de nature mathématique (au sens abstrait actuel).

2.4.2. **D'autre part, la Statistique est un « pont » entre théorie scientifique et observation phénoménale**.

On doit donc remarquer que, comme les Mathématiques, la Statistique a une origine liée à l'analyse de problèmes concrets.

Cependant, les « méthodes statistiques » relevant de certaines situations ne requièrent pas nécessairement la mise en oeuvre d'instruments sophistiqués : ainsi en est-il de l'estimation du coefficient de dilatation linéaire d'une barre métallique, ou du choix d'un instrument de mesure (horloge, thermomètre) dans un magasin. Une démarche statistique « inconsciente » permet donc de résoudre des problèmes courants : Mr JOURDAIN faisait aussi de la Statistique !

Dans certains contextes exigeant de la rigueur, il n'est pas toujours nécessaire de connaître le substrat statistique des méthodes utilisées. Ainsi :

(a) (physique) un ingénieur peut utiliser des tables de résistance des matériaux (étirement ou traction, pression, flexion, torsion, etc) sans connaître le mode d'élaboration de ces tables. Ces dernières peuvent notamment résulter d'expériences statistiquement contrôlées, dans lesquelles divers « facteurs » (types de matériaux, dimensions, volumes, masses, pressions, températures, etc) varient et influencent les capacités de résistance ;

(a) de même (biologie), un « praticien » utilise les résultats d'une analyse « biologique » (composants chimiques) ou physiologique (EEG, ECG, etc) issus de calculs statistiques (tables statistiques). Les « fourchettes » qui s'en déduisent servent généralement de repères et d'aides au diagnostic, à la thérapie ou au pronostic : cependant, elles ne supposent pas nécessairement la connaissance des techniques statistiques mises en oeuvre pour élaborer ces tables.

3. Les considérations précédentes, généralement peu connues du public, ont guidé la ***conception d'ensemble*** de ce dictionnaire. On raisonne à la fois **en largeur** et **en profondeur**.

3.1. Ce travail ne se limite pas à l'exposé de notions générales, comme c'est le cas des manuels de base. Au contraire, un **large éventail des outils existant dans la panoplie du statisticien** a été ouvert afin de couvrir une *grande variété de situations concrètes* : en effet, chaque situation peut être abordée de façon efficace à l'aide d'instruments figurant dans cette panoplie.

On peut souhaiter établir un jour un « ***mode d'emploi général*** » dans lequel chaque type de problème rencontré, une fois décrit et résumé à l'aide de « mots-clefs », se verrait associer un choix de démarches statistiques possibles (méthodes préconisées).

Pour ébaucher un mouvement dans ce sens, on propose une **aide au traitement de problèmes statistiques**. Celle-ci est très modeste, voire symbolique, mais elle devrait contribuer, dans certains cas, à faciliter le rassemblement des divers éléments (concepts, méthodes ou résultats) propres à traiter un problème scientifique donné. Elle gagnerait donc à être développée.

3.2. Par ailleurs, la maîtrise d'un **nombre limité de notions de base** (***information***, ***exhaustivité***, ***liberté***, ***invariance***), jointe à quelques **distinctions élémentaires** (***donnée*** et inconnue, ***théorique, empirique***, ***paramètre*** et ***observation***, ***observabilité*** et ***inobservabilité***), permet d'aborder d'emblée une large gamme de questions fondamentales : ***décision*** et ***optimalité*** (cf ***optimisation***), ***plan d'expérience***, ***sondage***, ***estimation***, ***prévision***, ***test d'hypothèses***, ***classification***, etc.

L'**apport unificateur de la théorie de la décision statistique** justifie la place centrale qu'elle occupe ici.

On peut, au passage, rendre hommage à un très grand statisticien, **Ronald A. Fisher** qui, pendant la première moitié du XX^{ème} siècle, a mis en évidence la plupart des concepts statistiques fondamentaux : ancillarité (cf ***liberté***, ***statistique ancillaire***), ***exhaustivité***, ***invariance***, ***vraisemblance***, expérimentation rationnelle, « ***randomisation*** », ***théorie fiducielle***, etc, largement exploités et développés depuis.

3.3. A l'inverse, un **problème de taille** était à résoudre : en effet, une compilation de ce genre ne peut raisonnablement prétendre traiter de toute la Statistique, déjà au seul niveau des théories, et encore moins à celui des applications qui sont extrêmement foisonnantes. S'il paraît utile d'approfondir différentes parties de la Statistique, on ne peut se perdre dans les labyrinthes de méthodes sophistiquées, ni incorporer des notions éphémères (eg créées pour les seuls besoins d'une démonstration).

Le contenu et l'agencement de ce site sont les résultantes de ces arbitrages, lesquels comportent toujours une part ... d'arbitraire.

4. Malgré la difficulté, on peut chercher à résumer l'**ensemble des critères retenus** pour circonscrire le champ à couvrir.

Une oeuvre de cette nature ne peut que constituer un ***compromis*** entre des objectifs ou des exigences multiples (cf avant-propos) et des nécessités pratiques de (dé)limitation. **L'orientation générale étant d'élaborer une oeuvre à la fois scientifique et multidisciplinaire**, sa conception a donc suivi les orientations suivantes.

4.1. D'abord, et ceci le différencie d'autres ouvrages, on a souhaité rassembler des ***définitions théoriques et rigoureuses des concepts et des méthodes statistiques*** sans, cependant, tomber dans l'abstraction pure. Ces définitions sont, autant que possible, données sous forme concise et suffisamment générale pour couvrir la plupart des situations rencontrées en pratique. Elles ne requièrent donc pas de connaissances, et surtout de développements, qui se situeraient au niveau de la recherche (troisième cycle universitaire). Ceci est un point de vue délibéré.

En effet, afin d'être accessible au plus grand nombre, la définition des notions, la description des méthodes ou l'explicitation des résultats ne sont pas toujours données dans leur plus grande généralité (abstraction). La contrainte majeure est de ne pas définir, ou de décrire, à une endroit une notion ou une méthode utilisées ailleurs avec un niveau de généralité (ou d'abstraction) supérieur.

L'intégration d'un « environnement » mathématique, décrivant ce que l'on appelle ici les ***méthodes mathématiques de la Statistique***, peut constituer une aide utile : algèbre (notamment algèbre linéaire), analyse (topologie, théorie de la mesure et de l'intégration, optimisation, etc), géométrie élémentaire. Ces méthodes sont les plus courantes en Statistique.

4.2. D'autre part, comme déjà souligné, ***le texte proprement statistique (donc mathématique)***, dont le formalisme nécessaire n'est pas à justifier, ***est généralement assorti de commentaires littéraires dont le but est pédagogique***. Le système des renvois (liens hypertextes) doit aider à trouver les notions cherchées ou leurs développements.

Ainsi, sans trop sacrifier aux exigences de rigueur, assorties d'un discours formalisé auto-suffisant, de nombreux développements et explications « littéraires » sont largement répartis dans le corps des articles.

Les articles purement mathématiques possèdent, par leur formalisme, l'avantage de fixer les notations utilisées, mais nécessitent en général peu d'explicitations.

4.3. Cette compilation est sous-tendue par un ***double effort de synthèse***.

4.3.1. Une première synthèse existe d'abord ***au niveau de chaque article*** : on a cherché à y faire le point des connaissances, au moins élémentaires, et d'une façon aussi concise que possible.

4.3.2. La deuxième synthèse, implicite, se situe ***au niveau d'ensemble*** : on a cherché à présenter une vue, à la fois large et approfondie, des différents « corps constitués » (ou « branches ») de la Statistique.

Tous les résultats produits par la recherche Statistique (passée et actuelle) ne peuvent évidemment être reportés. Le lecteur peut, le cas échéant, compléter le contenu de certains articles en fonction de ses centres d'intérêt ou de ses nécessités : les articles en question peuvent alors jouer le rôle de point de départ pour la rédaction de nouvelles « ***fiches de travail*** ».

Ceci explique pourquoi nous ne sommes pas allés trop loin dans les développements. En effet :

(a) d'une part *les **notions complexes** sont généralement constituées par association de notions simples*, donc plus accessibles. De plus, lorsqu'on se réfère à eux, les textes spécialisés définissent, si besoin, ces notions complexes. L'aide sommaire proposée pour le traitement de problèmes statistiques (cf supra § 3.1.) illustre ce phénomène d'agglutination ;

(b) d'autre part, *l'**intelligibilité du site** aurait été affaiblie* s'il avait été conçu selon une structure d'ensemble trop détaillée. En effet, une structure linéaire est plus simple qu'une arborescence, laquelle est plus simple qu'une structure en réseau. On a donc dû limiter le système des renvois, car un celui-ci devient très vite combinatoire, donc de plus en plus difficile à gérer.

Afin d'assurer la plus grande ***pérennité*** possible, un tri a été tenté entre des apports supposés éphémères ou secondaires dans l'abondante littérature spécialisée contemporaine, et des apports qui, par leur pertinence ou leur originalité, sont susceptibles de développements futurs d'intérêt général ou d'utilité pratique reconnue ou anticipable. Ces choix ont aussi tenu compte de l'importance (relative) du rôle joué actuellement par la Statistique dans chaque domaine de connaissance. Ici aussi, une part d'arbitraire est inévitable.

Les articles ont ainsi, en général, une étendue allant de quelques lignes à quelques « pages ».

4.4. Ce travail comporte des ***synthèses résolument originales***.

Un dictionnaire ne saurait être une oeuvre d'auteur. En effet, une compilation de ce type lui confère ipso facto une sorte d'indépendance ou d'« universalité », au moins dans son contenu : *les matières traitées sont toutes du domaine public de la Statistique*.

Néanmoins, l'effort de synthèse qu'il implique, ainsi que la réflexion qu'il a suscitée, ont conduit, dans plusieurs cas, à rédiger des ***présentations inédites et détaillées*** sur des thèmes qui sont souvent rapidement décrits, voire ignorés, dans les manuels.

C'est ce qui explique, et justifie, l'adoption du titre « Dictionnaire de la Statistique et de la Science ». Ainsi, divers ***articles fondamentaux*** exposent :

(a) les notions de ***domaine de connaissance***, de ***phénomène***, de ***système***, de ***structure*** ;

(b) des concepts relatifs à l'organisation du travail scientifique et statistique : ***observation***, ***système statistique***, ***production statistique***, ***consommation statistique***, ***situation statistique*** ;

(c) des méthodes ou théories de base du calcul des probabilités et de la Statistique : **Statistique descriptive**, ***théorie des processus***, ***théorie des plans d'expérience***, ***théorie des sondages*** ;

(d) diverses notions importantes : ***centralité*** et ***dispersion*** (ou ***variabilité***), ***forme***, ***conditionnement*** et **relation fonctionnelle** (cf aussi ***régression***, ***modèle de régression***), ***robustesse***, etc.

On doit alors aussi souligner l'importance de deux concepts fondamentaux en Science et en Statistique car situés au coeur-même de leur démarche :

(a) ceux de **domaine de connaissance** et de **phénomène** ;

(b) ceux de **loi multivariée** et de **relation fonctionnelle**.

On a explicité le contenu « intuitif » des notions présentées et, le cas échéant, cherché à montrer comment elles sont naturellement apparues dans le but de combler divers besoins, concrets aussi bien que théoriques.

5. On doit aussi préciser ***ce qui ne figure pas, ou n'est qu'effleuré, dans ce dictionnaire***.

5.1. En premier lieu, les ***notions mathématiques élémentaires*** (du niveau de l'enseignement secondaire) ne sont pas, sauf exceptions, incorporées au texte. D'autres dictionnaires, déjà anciens, traitent ces sujets : le « Dictionnaire raisonné de mathématiques » de André WARUSFEL (Le Seuil, 1966), le « Dictionnaire de Mathématiques » de Lucien CHAMBADAL (Hachette, 1981), ou encore celui de François LE LIONNAIS et collaborateurs (Presses Universitaires de France, 1979). Divers sites Internet (communautaires ou universitaires) contiennent une matière importante portant sur ces sujets.

5.2. Les notions liées à des domaines tels que la théorie des ***indices de prix*** (nombres indices) ou la ***démographie*** sont résolument ignorées. Ces domaines, parfois présents dans les manuels de Statistique (généralement assez anciens), comportent un intérêt et des difficultés propres. Mais leur nature n'est pas spécifiquement statistique, au sens entendu ici, mais sociologique (ie démographique ou économique). Lorsqu'ils font référence à des lois spéciales (***fonction de survie***, lois de mortalité, etc), ou à des caractéristiques telles qu'indices, rapports, taux ou quotients, ces domaines comportent alors des aspects statistiques, mais il s'agit d'applications de la Statistique, même si les domaines considérés ont pu avoir été à l'origine de concepts premiers.

En matière de démographie, on peut se référer eg au « Dictionnaire de démographie » de Roland PRESSAT (Presses Universitaires de France, 1979), ainsi qu'au « Dictionnaire de démographie et des sciences de la population » (oeuvre collective sous la direction de France MESLÉ, Laurent TOULEMON et Jacques VÉRON, édition spéciale de l'INED, librairie Armand COLIN, Paris, 2011)**.**

5.3. Pour des raisons voisines, on ne fait qu'effleurer le ***contrôle de réception*** (contrôle de la qualité des productions). Il s'agit d'un domaine d'applications spécialisées dont les techniques (***carte de contrôle***, etc) n'apportent guère à la Statistique mais, au contraire, lui doivent certaines solutions (***test d'hypothèses***, ***théorie séquentielle***, etc). Il en va de même pour ce qu'il est convenu d'appeler la ***recherche opérationnelle***.

Sur ces sujets, divers organismes, comme l'Association française de normalisation (A.F.NOR.) (**http://www.afnor.org/**), ont publié des documentations spécifiques. Cette matière relève aussi bien de la normalisation que de la règle de droit.

5.4. N'étant pas particulièrement ***orienté*** vers les applications mais plutôt ***vers la compréhension de la Statistique***, ce travail ne fournit guère beaucoup de détails sur les questions suivantes :

(a) **imagerie** et **visualisation**. Le rôle des ***représentations graphiques*** est très important en Statistique, notamment dans les applications. De nombreuses techniques ont été développées à cette fin ; ces techniques ont été améliorées grâce à l'évolution des fonctionnalités de l'outils informatique. Plusieurs graphiques (à caractère « théorique ») ont été intégrés lorsqu'ils semblaient utiles pour la pédagogie d'un concept, d'une méthode ou d'un résultat. A défaut (formes des familles de densités, etc), les informations nécessaires à l'élaboration des graphiques sont, en général, explicitées de façon à en suggérer et faciliter le tracé ;

(b) **tables statistiques**. Cette limitation résulte de considérations de volume évidentes. Certaines revues spécialisées contiennent des programmes informatiques, parfois sous forme de codes sources, associés à des méthodes publiées : calculs de nombres au hasard, tabulations de lois particulières, etc.

Les **tables des *lois de probabilité*** et de leurs caractéristiques empiriques figurent eg dans le « Handbook of mathematical functions, with formulas, graphs and mathematical tables » (notamment, Chapitre 26), par Milton ABRAMOVITZ et Irene A. STEGUN, Editors, décembre 1972 (Dover Publications, Inc., New York).

Les familles de lois de probabilité utilisées en Statistique sont très variées. Cependant, on pourra noter que les ***propriétés asymptotiques*** énoncées dans ce dictionnaire permettent souvent de se ramener à des tables usuelles, notamment à celles des lois de probabilité les plus simples (cf ***approximation des lois***) ;

(c) les **méthodes de calcul numérique** mises en oeuvre en Statistique. Elles ont leurs difficultés propres, mais les logiciels informatiques actuels utilisent des procédures de calcul performantes, dont ils décrivent généralement les principales propriétés numériques (précision, nombre d'itérations, etc). On peut cependant observer la persistance d'une course-poursuite entre l'accroissement de la puissance de calcul et l'accroissement de consommation dûe à des besoins de plus en plus sophistiqués. L'augmentation de puissance permet, en effet, d'atteindre et de traiter des situations antérieurement inaccessibles : **nombres au hasard**, **sondages**, **simulations**, **grandes bases de données** (GBD).

5.5. Enfin, ***les vieilles tournures*** (eg « *résumé exhaustif* ») ***ainsi que les jargons et anglicismes*** qui dénaturent la langue française sans nécessairement rendre l'exposé plus clair, ont été exclus.

Pour limiter cet inconvénient, notamment lorsqu'une expression est attachée à des noms d'auteurs, on l'associe simplement à ces derniers. Ainsi, certains anglicismes notoires ont parfois été conservés et explicités lorsqu'ils ne comportaient pas d'équivalent français, eg :

* « studentisation » peut s'exprimer par « ***transformation de STUDENT*** » ;
* « winsorisation » peut s'exprimer par « ***transformation de WINSOR*** » ;
* « bootstrap » peut s'exprimer par « ***méthode de EFRON*** » ;
* « Jack-knife » (ou « jacknife ») devient « ***méthode de QUENOUILLE*** » (ou de QUENOUILLE-TUKEY).

5.6. Un dictionnaire est un ***instrument de travail***, et non pas une encyclopédie de culture générale.

Les aspects biographiques ou historiques n'y figurent donc pas sauf, implicitement, à travers des noms d'auteurs, lorsque ceux-ci sont restés attachés à un concept important, à une méthode de base ou à un résultat fondamental. Aucune recherche de paternité n'a été tentée : c'est l'usage qui a globalement prévalu. Il arrive souvent que plusieurs auteurs soient à l'origine d'une même notion, ou aient contribué de façon significative à l'extension, à la généralisation ou à l'étude approfondie des propriétés relatives à cette notion : dans ces cas, leurs noms apparaissent simplement et systématiquement dans l'ordre alphabétique. Ainsi, l'inégalité dite « de FRÉCHET-DARMOIS » en France et « de CRAMER-RAO » dans les pays anglo-saxons, a été appelée « ***inégalité de CRAMER-DARMOIS-FRÉCHET-RAO*** ».

5.7. Des raisons allant dans le même sens ont aussi entraîné l'***exclusion de toute référence bibliographique***. Des allusions peuvent être perceptibles. Elles apparaissent implicitement, dans l'***index des auteurs*** :

(a) à travers des dates : dates biologiques (naissance, décès) et dates fonctionnelles (principales publications) des auteurs ;

(b) ou encore à travers leurs « domaines d'intervention » : principaux types de travaux, classifiés par thèmes selon la ***classification thématique générale***.

6. Il existe actuellement (2020) un certain nombre de documents appelés ***dictionnaire de Statistique*** ou ***encyclopédie de Statistique*** (cf supra). Ils se présentaient traditionnellement sous forme d'ouvrages mais, progressivement, des sites Internet proposent en ligne des contenus équivalents. Ainsi en est-il des documents suivants.

6.1. Ouvrages en langue française :

* « Dictionnaire de Statistique » de Eugène MORICE (Editions DUNOD, 1968) ;

* « Glossaire systématique quadrilingue (anglais-français-espagnol-russe) de la terminologie des méthodes statistiques », élaboré par Isaac PAENSON et ses collaborateurs (1970) et publié sous l'égide de l'ONU ;

* « Recueil de normes de la Statistique », édité par l'AFNOR (deuxième édition, 1978).

6.2. Ouvrages en langue anglaise

Sans être exhaustive, la liste ci-dessous peut se révéler utile :

* « Dictionary of statistical terms » de Maurice George KENDALL et William R. BUCKLAND (cinquième édition par F.H.C. MARRIOTT, 1990 ; sixième édition par Yadolah DODGE, 2003). Cet ouvrage est publié depuis de nombreuses années pour le compte de l'Institut International de Statistique (IIS) ;

* International encyclopaedia of Statistics (Encyclopédie internationale de la Statistique), éditée par William KRUSKAL (1978) ;

* Encyclopaedia of statistical sciences (Encyclopédie des sciences statistiques) (9 tomes), éditée par Samuel KOTZ et Norman L. JOHNSON, avec Campbell B. Read (1981 - 1989), chez Wiley Interscience ;

* « **Concise statistical vocabulary** : english - french - chinese » (précis de vocabulaire statistique anglais - français - chinois), par Véronique Alexandre et Jean Alain Monfort, sous la direction de Pascal Mazodier, publié par le State statistical bureau (SSB), Beijing, 1998
(cf aussi http://bibliotheque.insee.net/index.php?lvl=author_see&id=39855)

6.3. Sur Internet

La liste ci-après ne prétend pas, non plus, à l'exhaustivité :

* sites universitaires ou de grandes écoles (accès à des pages de statistique : notes de cours, etc), eg :

** Lyon 2 :
(**http://eric.univ-lyon2.fr/~ricco/cours/cours/Probabilites_et_Statistique.pdf**) ;
** INSA de Lyon (**http://rfv.insa-lyon.fr/~jolion/STAT/poly.html**) ;
** Paris 5 Descartes (**http://w3.mi.parisdescartes.fr/smel/cours/sd/sd.html**) ;
** CNAM de Montpellier :
(**http://www.agro-montpellier.fr/cnam-lr/statnet/cours.htm**) ;

* Wikipedia, site à vocation encyclopédique, contient :

** des pages de probabilités (**http://fr.wikipedia.org/wiki/Probabilit%C3%A9**) ;
** des pages de statistique (**http://fr.wikipedia.org/wiki/Statistiques**).

L'intérêt des **supports d'information** précédents en matière de terminologie est incontestable : ils contribuent progressivement à établir une certaine ***« normalisation » du vocabulaire statistique***, parfois encore « flottant ». Cependant, on peut regretter le caractère « scolaire » ou « partiel » de certaines présentations, qui tendent à réduire la Statistique à des fonctions seulement calculatoires ou procédurales (cf Avant-propos).

Les objectifs recherchés dans les textes précédents ne coïncident donc pas nécessairement avec ceux du présent dictionnaire. En outre, les encyclopédies ont souvent l'inconvénient de ne pas être de mise en pratique immédiate (inclusion de biographies, etc) ou d'inclure (notamment lorsque divers contributeurs interviennent) des redites ou des doubles emplois (eg références bibliographiques redondantes), ainsi que des systèmes de notation différents qui rendent moins commodes les transitions d'un article à un autre.

7. Un dictionnaire peut a priori comporter ***plusieurs organisations possibles***.

7.1. Son **architecture** peut s'inspirer des idées directrices suivantes :

(a) au sens « étroit », une ***structure sans liens*** comporte des articles présentés de façon indépendante les uns des autres, alors même que des liens (logiques ou autres) existent ;

(b) au sens le plus large, une ***structure en réseau*** représente la forme la plus générale, et aussi la plus neutre, d'organisation. Elle consiste à relier toutes les notions statistiques, même, le cas échéant, lorsque certains liens n'ont guère de pertinence (réseau global « en aveugle »).

Cette deuxième structure manque évidemment de pertinence, puisque seul un sous-graphe du réseau correspond à des relations intéressantes. Mais elle est évoquée ici pour situer la question dans un contexte global. De plus, compte tenu de l'intérêt des association d'idées, cette structure peut préfigurer des liens à venir entre notions statistiques actuellement non reliées ;

(c) entre les deux extrêmes précédents, on peut concevoir des situations intermédiaires. Ainsi :

** une ***structure de type linéaire*** présente, indépendamment les unes à la suite des autres, des notions statistiques sans préciser de façon détaillée les liens entre elles. Ainsi, chaque article renvoie à un seul autre. Les inconvénients de cette structure sont évidents ;

** une ***structure de type emboîté*** (arborescence), propice à une démarche thématique, inclut, sous un thème donné, l'ensemble des notions se rapportant à ce thème. Son défaut vient des imputations multiples, un même article relevant de plusieurs thèmes ;

7.2. Finalement, une solution moyenne, de type mixte, a été suivie. Elle adopte une **structure en réseau « raisonnée »** : on y trouve des liens privilégiés (thèmes) lorsque les articles correspondent à une même « branche » de la Statistique, ou une absence de lien lorsque les articles se suffisent à eux-mêmes (outils mathématiques de la Statistique, ou articles banalisés en raison de leur polyvalence). Ainsi, ce dictionnaire incorpore :

(a) des ***articles généraux relatifs à des théories ou à des méthodes***, lesquels présentent, de façon succincte mais structurée, la plupart des notions (concepts ou expressions conceptuelles, méthodes ou procédures, principaux résultats) qui se rapportent aux articles en question ;

(b) des ***articles plus spécialisé***s, ou à caractère technique, conçus en vue de préciser certaines notions (concepts ou méthodes), les principales propriétés qui leur sont attachées (ie pour lesquelles ces notions ont été conçues) ainsi que les notations en usage.

Les inconvénients possibles de cette architecture sont cependant limités grâce à l'usage des **liens hypertextes**, qui facilitent la navigation interne. Le lecteur a d'emblée une vue générale qui lui permet de situer, grâce au système des renvois, les principaux corrélats associés à ces articles ou à ces notions. Ces corrélats permettent, si besoin, d'approfondir les notions étudiées.

8. ***Aucune démonstration*** (au sens mathématique du terme) ne peut tenir de place dans un dictionnaire. En inclure aurait augmenté sensiblement sa taille et réduit sa

lisibilité. De plus, celà reviendrait à un mélange de genres : dictionnaire vs manuel. En contrepartie, on a dû observer un soin scrupuleux dans la vérification des formules, afin de limiter les risques d'erreurs (symboles utilisés, formules reproduites).

9. Un mot doit être dit sur l'important ***problème du symbolisme***. Etant donné la variété des situations décrites, et malgré le lien unificateur de la ***théorie de la décision*** (statistique), il est impossible d'adopter une ***notation uniforme de bout en bout***. Sauf à adopter un système de notations complexe, voire compliqué (augmentation du nombre de symboles, augmentation des artéfacts : indices, exposants, groupages, etc), des incohérence s'introduisent rapidement (cf infra, § 9.2.).

9.1. Pour les ***notations plus ou moins consacrées par l'usage***, ou ***notations « standards »***, les symboles usuels sont conservés dans l'ensemble des articles, quitte à en indiquer des « variantes » dans d'autres articles. Plusieurs thèmes (ou « branches ») - pour ne citer que les principaux abordés en calcul des probabilités et en Statistique - se prêtent assez bien au maintien d'un symbolisme homogène : ***théorie des processus***, ***théorie de la décision***, ***théorie de l'estimation***, ***théorie des tests***, ***théorie des sondages***, ***théorie des plans d'expérience*** ;

9.2. Dans d'autres domaines, ***les notations ont été rendues aussi uniformes que possible***.

Mais cet exercice rencontre vite des **limites** (incohérences formelles). Ainsi :

(a) lorsqu'un indice i varie, par valeurs entières, de 1 à n dans une définition, et que celle-ci renvoie à un article dans lequel on utilise le symbole n comme indice variant par valeurs entières de 1 à N, le lecteur devra ***effectuer les transcriptions*** ad hoc, soit mentalement dans les cas simples, soit par écrit, à l'aide d'une table de passage entre les deux systèmes de notations, dans des cas plus complexes. C'est d'ailleurs cette démarche qui a été suivie ici pour élaborer des articles de synthèse résultant de plusieurs écrits (articles de revues, manuels) ;

(b) lorsque T désigne l'ensemble des temps d'un processus ou d'une série temporelle, il arrive que cet ensemble soit écrit {1 ,..., T} (cas discret) ou [0,T] (cas continu) : autrement dit, on pourra noter à l'aide du ***même symbole*** l'ensemble des temps et sa « borne » supérieure ;

(c) de même, la notation adoptée pour un ***processus*** $\{(\Omega, \mathcal{T}, P), (\mathcal{X}, \mathcal{B}), (X_t)_{t \in T}\}$ conduit à utiliser la lettre ω pour désigner un élément de Ω (ou une trajectoire particulière) ; or, cette lettre peut aussi désigner (eg en ***analyse spectrale***) une ***fréquence***. Dans de tels cas, cependant, le contexte permet de lever la difficulté.

Ce type d'inconvénient, que nous avons essayé de réduire, semble compensé par le fait qu'une notion est ainsi rendue autonome, en termes de notations, et qu'elle peut donc être utilisée dans des contextes différents (ceci est notamment le cas des articles banalisés, relatifs aux « méthodes mathématiques de la Statistique ») :

l'inconvénient de la « traduction » est balancé par la « polyvalence notationnelle » des définitions.

Finalement, la plupart des notations adoptées, parmi les plus répandues, permet de couvrir une large gamme de situations, telles que celles exposées ici. Le lecteur prévenu ne devrait guère rencontrer d'ambiguïté qu'il ne sache résoudre aisément. L'***index des principales notations*** sert aussi de guide de réflexion.

10. D'un point de vue pratique, l'***utilisation du dictionnaire*** peut s'effectuer à l'aide des outils suivants :

(a) des ***index analytiques alphabétiques*** : ceux-ci renvoient aux articles, à raison d'un index par lettre de l'alphabet (avec regroupement des lettres « rares »). On peut accéder à la plupart des notions par le biais de ces index et de leurs liens hypertextes ;

(b) un ***index des principales notations*** ;

(c) un ***index des noms d'auteurs*** (avec imputation thématique et datation) ;

(d) une ***classification thématique générale***, élaborée à partir des principales classifications existantes :

* celle de l'American Statistical Association ;
* celle de l'Institut International de Statistique (Statistical theory and Method abstracts, vol. 31, n° 4, 1990) ;
* celle de la revue « Communications in Statistics, Theory and Methods » (Ed. in chief D.B. OWEN, chez Marcel DEKKER, Inc.).

*

* *

*

Nées, au départ, de diverses *notes de lecture accumulées* par l'auteur depuis le début des années 1980, le matériel ainsi rassemblé a été organisé en cherchant à faire preuve à la fois d'un **esprit de système** (pour les questions de synthèse et de normalisation), d'un **esprit de pertinence** (pour les choix effectués) et d'un **esprit d'organisation** (pour la présentation adoptée). Diverses personnes (collègues, étudiants, professionnels divers) intéressées par cette démarche, ont aidé à améliorer cette organisation.

Toutes suggestions susceptibles d'augmenter encore la qualité de ce site seront aussi les bienvenues.

Saint Georges de Monclar
1980-2020

RETOUR A L'ACCUEIL **VERSION ANGLAISE**