

YANNICK MACÉ

## **L'approche statistique : entre réalité(s) et subjectivité**

*Journal de la société française de statistique*, tome 147, n° 4 (2006),  
p. 85-102

[http://www.numdam.org/item?id=JSFS\\_2006\\_\\_147\\_4\\_85\\_0](http://www.numdam.org/item?id=JSFS_2006__147_4_85_0)

© Société française de statistique, 2006, tous droits réservés.

L'accès aux archives de la revue « Journal de la société française de statistique » (<http://publications-sfds.math.cnrs.fr/index.php/J-SFdS>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme  
Numérisation de documents anciens mathématiques  
<http://www.numdam.org/>

# L'APPROCHE STATISTIQUE : ENTRE RÉALITÉ(S) ET SUBJECTIVITÉ

Yannick MACÉ\*

« Il est notoire qu'il n'existe pas de classification de l'univers  
qui ne soit arbitraire et conjecturale. La raison en est fort simple,  
nous ne savons pas ce qu'est l'univers. »

J.L. Borges.

*Autres inquisitions.*

## RÉSUMÉ

Pour penser, l'être humain doit percevoir et saisir le sensible qui est, par nature, infini. L'homme a ainsi développé un mécanisme de catégorisation, de classification du réel. Ce processus se retrouve dans l'essence de toute discipline liée à la pensée humaine, langage ou sciences par exemple. Un tel mécanisme est intrinsèquement lié à une perte d'information par rapport au « réel ». Les techniques statistiques entrent de manière évidente dans ce schéma. Et c'est, entre autres, dans cet espace de subjectivité que beaucoup de questions associées à la déontologie de notre métier trouvent leur place.

## ABSTRACT

In order to think, the human being needs to perceive and to grasp the physical world; by nature, that world is infinite. So mankind developed a mechanism of categorization, a classification of reality. Such a process lies in the very essence of language or science for example. This mechanism, however, bears in itself an inevitable loss of information respective to « reality ». Obviously, statistical techniques fall within this process. And, many ethical issues of our profession also.

## 1. Introduction

Une statistique peut-elle être fausse ? A contrario peut-elle être vraie ? Qu'est-ce que la réalité mesurée par le travail d'un statisticien ? Cette question ne renvoie-t-elle pas à la question de la perception de la réalité, d'une manière plus générale ?

Cet article propose la mise en forme de quelques idées qui montreraient que, comme tout processus de la pensée humaine, la nature même des approches

---

\* Groupama SA, 5-7 rue du Centre, 93799 Noisy-le-grand.  
yannick.mace@groupama.com

statistiques passe par une simplification incontournable du réel. Comme d'autres activités intellectuelles, les statistiques découpent le réel et le recomposent pour mieux le discerner. L'appréhension totale de notre environnement immédiat comme celle de notre univers, au sens large et matérialiste du terme, relèverait d'une perception continue de l'infini. Pour saisir cet environnement nous avons besoin de fragmenter et de compartimenter le monde qui nous entoure, de passer de l'infini au fini, de l'abstrait au concret, du concept à l'ordre et à la mesure. C'est une des caractéristiques universelles du mode de penser humain. La démarche statistique s'inscrit bien évidemment dans ce schéma. On touche ici à l'essence du métier de statisticien. Cette façon de voir les choses n'est pas sans conséquence sur le sens qu'il est possible de donner à notre profession, sur le caractère subjectif de nos travaux et sur l'engagement moral et déontologique qui en découle. Afin d'illustrer ces idées en sortant de notre cadre de formulation classique et des discussions épistémologiques des statisticiens eux-mêmes, le propos prendra plus souvent appui sur diverses citations d'auteurs a priori étrangers à notre profession.

## 2. Existence ou perception du réel ?

Le réel a-t-il une existence ? Quelle perception peut-on avoir de la réalité ? La notion de concept, au sens platonicien du terme, est-elle suffisante pour saisir l'intégralité de la réalité, la réalité au sens matérialiste du terme ?

Borges tout au long de son œuvre est resté passionné par le sujet. Dans sa fiction « Funes et la mémoire », qui met en scène un jeune prodige uruguayen, Borges traite l'argument en utilisant l'archétype de la feuille de l'arbre ou celui du chien : « En effet, non seulement Funes se rappelait chaque feuille de chaque arbre de chaque bois, mais chacune des fois qu'il l'avait vue ou imaginée ... Non seulement il lui était difficile de comprendre que le symbole générique *chien* embrassât tant d'individus dissemblables et de formes diverses ; cela le gênait que le chien de trois heures quatorze (vu de profil) eût le même nom que le chien de trois heures un quart (vu de face). » N'existe-il pas une infinité de feuilles ou de chiens possibles correspondant à chacun des états successifs de la nature ou à un angle de vue différent du sujet qui les observe ?

Toujours dans le texte déjà cité, « Funes ou la mémoire », Borges étend sa réflexion au plus loin afin de montrer que la mémoire ne peut être qu'associée au principe de sélectivité ; l'appréhension totale de l'univers conduirait à l'insomnie puis à la mort inéluctable. Selon lui, le salut pour l'esprit humain ou pour la mémoire humaine réside dans la simplification ou la capacité d'oublier, d'occulter, de filtrer une partie du réel. Funes qui ne disposait pas de cette faculté d'oubli apprenait sans effort différentes langues au fur et à mesure de ses lectures, concevait des systèmes de numération sans logique de construction (qui s'apparentaient alors à un empilement exhaustif de quantités), vivait une perception du temps exacte, précise et continue (toutes les informations qu'il percevait restaient accumulées dans sa mémoire). Le seul destin possible pour ce personnage de fiction qui mémorisait l'infini et qui, par conséquent, était incapable de penser, l'a alors conduit à l'insomnie

puis à sa fin prématurée. Borges de conclure en quelque sorte : « Penser, c'est oublier des différences, c'est généraliser. »

Percevoir le réel, dans toutes ses dimensions, dans sa globalité apparaît donc impossible à l'esprit humain dont les capacités sont limitées. D'où la caractérisation proposée par Edgar Morin : « le réel est monstrueux. Il est énorme, il est hors norme ». Pour survivre, pour penser, il faut pouvoir appréhender la réalité, pour l'appréhender il faut la simplifier : passer de l'infini au fini.

Ce processus inhérent à la pensée a conduit l'homme à découper, à fragmenter la réalité (ou la perception qu'il peut en avoir) et à construire naturellement des catégories, des ensembles, des sous-ensembles, des divisions, des subdivisions, des groupes, des sous-groupes, des typologies, des segments, des partitions, des familles, ...

Un tel mécanisme régit tous les domaines liés au fonctionnement de l'esprit humain, jusqu'à nos perceptions du quotidien. Georges Perec rappelle qu'il est : « Tellement tentant de vouloir distribuer le monde entier selon un code unique ; une loi universelle régirait l'ensemble des phénomènes : deux hémisphères, cinq continents, masculin et féminin, animal et végétal, singulier pluriel, droite gauche, quatre saisons, cinq sens, six voyelles, sept jours, douze mois, vingt-six lettres. »

La partie suivante évoque quelques assises philosophiques et exemples empruntés à l'histoire des sciences naturelles, pour appuyer l'idée du caractère universel de ce besoin de catégorisation du réel.

### 3. Classer pour percevoir, percevoir pour penser

Un léger détour vers l'étymologie nous rappelle qu'une analyse est une *décomposition* (*analysis* en grec ancien). Penser pour comprendre ; comprendre (*comprehendere* en latin classique) c'est *saisir ensemble, prendre ensemble* un certain nombre d'objets, de caractéristiques ou d'idées dont on met en évidence les propriétés communes en les inscrivant dans un concept. Celui-ci est formé par élimination de quelques détails, par schématisation, généralisation ou abstraction. Les objets isolés, divers et éparpillés étaient mal aisés à distinguer, groupés avec d'autres au sein d'une même catégorie, ils sont plus facilement identifiés.

Depuis l'approche platonicienne de la connaissance, le concept est considéré comme la notion centrale qui permet à la pensée de passer du multiple à l'unique, du pluriel au singulier, du différent au semblable. La finalité de l'intelligence consiste à se détourner de la réalité sensible, les images ne sont alors que les reflets des concepts. Le concept de carré se visualise certes par la figure géométrique d'un carré dessiné dans le sable ou au tableau, mais ne se réduit pas à ce seul dessin. L'essentiel de la pensée, le véritable réel est dans l'Idée, dans le concept (qui se définit pour cet exemple comme une figure géométrique dont les angles sont droits et les quatre côtés égaux).

Avec Descartes et la description de sa « méthode » en 1637, il nous est devenu familier de considérer que la nature de l'ensemble des sciences (la mathesis universalis, la mathématique universelle) est régie par un fondement commun et universel. Celui qui consiste à appréhender notre réalité matérielle au travers de notions et d'instruments qui apportent « ordre et mesure » pour aider notre raison à comprendre le monde qui nous entoure.

Un des (quatre) préceptes que Descartes se propose d'observer, pour bien conduire sa raison, n'est-il pas de « diviser chacune des difficultés que j'examinerais en autant de parcelles qu'il se pourrait, et qu'il serait requis pour les mieux résoudre » ? Puis de « ... conduire par ordre mes pensées, en commençant par les objets les plus simples et les plus aisés à connaître, pour monter peu à peu comme par degrés jusques à la connaissance des plus composés ... » ? Finalement, Descartes arrive à la conclusion selon laquelle « ... les choses que nous concevons fort clairement et fort distinctement sont toutes vraies. » Le processus de découpage du réel permet de clarifier la perception du monde et aide à sa compréhension pour mieux approcher la vérité.

Les sciences de la nature, de la chimie à la biologie se sont construites sur de nombreux types de classifications. Cela va du tableau de Mendeleïev qui établit la classification périodique des éléments (lequel permet de visualiser les différents composants chimiques de la matière), aux diverses nomenclatures des êtres vivants proposées par les naturalistes.

En ce sens, l'histoire des sciences naturelles est très révélatrice. Face à la diversité quasi infinie du vivant, l'objectif des naturalistes du XVII<sup>e</sup> au XIX<sup>e</sup> siècle était de faire l'inventaire et de procéder à la description de cette diversité. Cette quête a donné lieu au développement des techniques de classification : la taxinomie.

Si l'on remonte à l'antiquité, Aristote (*Historia animalium*), dans ses activités de naturaliste, avait déjà tenté d'élaborer une classification animale. Il avait abouti, selon un schéma hiérarchique à séparer les oiseaux des poissons, puis avait introduit une séparation entre les animaux « dotés en sang » et ceux « dépourvus de sang », puis entre ceux « dotés de quatre pattes », ceux « dotés de nombreuses pattes » et ceux « sans pattes ». Linné proposa une classification (des animaux et des plantes) en règne, classe, ordre, genre et espèce. Cuvier introduisit des éléments issus de l'étude de l'anatomie interne et sépara les vertébrés des invertébrés. Il affina ainsi la classification des invertébrés en précisant les ordres des mollusques, des méduses et anémones de mer. Puis dans la lignée de l'évolutionnisme de Darwin, la détermination des groupes se fonda sur l'existence d'ancêtres communs aux différentes espèces. Aujourd'hui le processus s'est encore précisé, distinguant par exemple les organismes unicellulaires chez les animaux et, en botanique, distinguant les champignons comme un règne distinct de celui des plantes. Il y a eu depuis l'antiquité une évolution constante des méthodes et des critères de classification utilisés. Cette évolution a été induite par les progrès techniques (invention des herbiers, invention du microscope, exploration de l'ADN, ...), par la découverte ininterrompue de nouvelles espèces d'êtres vivants et par les évolutions méthodologiques (le passage d'une méthode de divisions

par dichotomies successives – méthode descendante – à une méthode par regroupements successifs – méthode ascendante –).

À l'évidence, d'autres exemples empruntés aux sciences physiques, humaines, économiques, géographiques,... pourraient tout aussi bien le souligner, l'intelligence humaine pour se structurer et se développer, n'a d'autre issue que de se fonder sur un découpage du réel. La pensée se nourrit de perceptions, pour rendre ces dernières intelligibles, elle procède à une fragmentation du réel. Pour approcher et distinguer les choses, pour s'en faire une idée tangible, il faut s'en tenir à l'appréhension d'un nombre limité de caractéristiques distinguables.

#### **4. L'approche statistique dans ce schéma**

Étudier totalement une population (un ensemble de taille importante de données sur des individus) sans recours à la statistique reviendrait à vouloir percevoir de manière simultanée et continue, toutes les caractéristiques de tous les individus.

La démarche statistique permet de passer de l'observation de cette population de départ (au sein de laquelle chaque individu la composant est de mesure nulle) à une approche par individu «statistique». C'est ainsi qu'elle permet de mesurer les proximités entre caractéristiques, de construire des typologies ou de tester des hypothèses. En définitive, elle permet de donner un contenu quantitatif aux concepts qu'ils soient de nature sociologique, économique, démographique ou autre. Devant un ensemble de données, les procédures et indicateurs statistiques ont bien vocation à proposer des filtrations de la réalité, afin de décrire les populations étudiées, de tenter de réaliser des prédictions ou de concevoir des outils d'aide à la décision.

Un rapide panorama, qui ne se voudra pas exhaustif, des techniques les plus couramment utilisées à ce jour, confirmera que cet objectif de simplification est bien au cœur de la pratique statistique.

##### **i) Les statistiques descriptives**

L'ensemble des indicateurs de statistiques descriptives tels que les répartitions d'individus sur une variable discrète, les indicateurs de médiane ou de moyenne, les indicateurs de forme des répartitions (variances, écarts types, coefficients de variation, coefficients d'asymétrie, coefficients d'aplatissement ou quantiles), les ajustements de distribution d'une quantité à une loi ou à une famille de lois ou les coefficients de corrélation, entrent dans ce cadre d'examen. Ils peuvent être vus comme des simplifications extrêmes; un ou plusieurs nombres synthétisent l'ensemble de l'information portant sur une population.

La perte d'information associée à l'utilisation de statistiques descriptives provient de l'écart entre les données de départ et leur représentation simplifiée par ces indicateurs qui mettent en lumière quelques-unes seulement de leurs caractéristiques.

## ii) L'analyse des données

Les techniques dénommées couramment « analyse des données » entrent tout à fait dans ce schéma, qu'elles servent à élaborer des segmentations, techniques de découpage d'une population, ou aient pour objet de révéler des formes, des agencements cachés tout en ayant la capacité de mettre en évidence les éléments extrêmes ou atypiques.

Les analyses en composantes principales (incluant l'analyse des correspondances et divers autres avatars) sont des techniques fondées sur la projection de grandes quantités de données sur des espaces de dimensions réduites, de manière à déceler les relations entre grandes familles de caractéristiques. Les proximités recherchées peuvent porter aussi bien sur l'univers des individus que sur celui des variables étudiées. Les projections se réalisent sous contrôle de perte d'information – ou « perte d'inertie » – minimum. Les axes retenus qui servent à visualiser les projections sont ceux dont on considère qu'ils apportent le plus d'information.

Dans cette même veine, les techniques de classifications (souvent associées aux analyses ci-dessus) ont pour objet de regrouper les individus d'une population de taille importante au départ, au sein de groupes les plus homogènes possibles (sous contrainte d'être les plus hétérogènes possibles entre eux). Que ce soient les méthodes hiérarchiques (ascendantes ou descendantes), les méthodes de partition (de type « nuées dynamiques ») ou celles de classification floue, leur caractère efficient repose sur la minimisation d'une distance entre les données de départ et leur représentation sous forme de classes. Il y a bien au cœur de ces techniques la recherche de maximisation de la quantité d'information « utile » révélée pour l'analyse. Et toutes reposent sur des notions analogues de mesure de distance entre partitions successives : indices de mesure de la similarité entre individus, coefficient d'affinité, coefficient de séparabilité, mesure de concordance, critère de perte d'inertie minimale, critère du saut minimal. Ces mesures ont à charge d'évaluer l'éloignement entre la population de départ et sa représentation sous forme « classée » ou « projetée ».

## iii) Les techniques de modélisation en économétrie

L'économétrie constitue une famille de techniques dont il est possible de dire, sous forme imagée, qu'elle permet d'ajuster une tendance autour d'un nuage de points. Les modèles économétriques sont ainsi typiquement des instruments simplificateurs. Les modélisations, qu'elles aient un rôle prospectif ou explicatif, qu'elles portent sur des données déjà agrégées (« macro ») ou sur des données individuelles (« micro ») ont toujours pour objet de lier un phénomène à quelques variables et de trier parmi un ensemble de variables explicatives celles qui se révèlent les plus pertinentes. Leur finalité, d'une manière générale, consiste à dégager des régularités, des tendances (pour les séries chronologiques indexées sur une dimension temporelle) qui permettent une fois dessinées de faciliter l'interprétation, l'explication du présent ou du passé, la prévision ou la simulation.

En définitive, la démarche afférente aux modélisations propose une simplification du réel d'une part par la formalisation de relations (linéaires pour les

plus simples d'entre elles) et d'autre part par l'utilisation des lois du hasard. Or, dans une conception déterministe du monde, la justification de l'utilisation des lois du hasard provient, non de ce que les phénomènes observés sont régis par des lois purement aléatoires mais plutôt de ce que ces lois, en simplifiant les approches, rendent compte efficacement d'une réalité infiniment complexe. Il y a, au centre de ces méthodes ce statut épistémologique particulier de la notion de probabilité. De manière schématique, on pourrait affirmer que la mise en œuvre de modèles aboutit à une réduction de  $N$  équations ( $N$  quasi infini, du moins par rapport aux capacités de l'esprit humain, au traitement possible par les ressources informatiques et à la possibilité de les alimenter en données) à quelques équations stochastiques. Lesquelles permettent alors la description, le raisonnement et la prévision. Il n'y a donc pas contradiction entre la perception de phénomènes apparemment aléatoires et leur réalité déterministe. Aussi, les notions de probabilités utilisées dans ce cadre relèvent plus d'un besoin de simplification et de formalisation du réel que de la mesure du caractère purement aléatoire d'un phénomène.

La recherche de l'ajustement optimum, l'ensemble des indicateurs d'écart entre la « vraie » loi inconnue et le modèle proposé ainsi que les tests d'adéquation doivent s'interpréter comme la quête de la meilleure appréhension du phénomène étudié. La perte d'information est ici due aux hypothèses de spécification du modèle et des lois sous-jacentes au regard d'une distribution des données qualifiée d'aléatoire.

#### iv) La théorie des sondages et les techniques d'enquête

La théorie des sondages fournit un cadre méthodologique du passage de l'échantillon à l'extrapolation sur une population exhaustive. Cette théorie peut s'identifier à un corpus de formules qui décrit et formalise la perte d'information liée à l'élaboration de résultats établis sur des sous populations. Elle nous expose de ce fait le mode optimal de constitution d'échantillons dans le but de diminuer la déperdition d'information.

Plus largement, les différentes étapes du sondage ou de l'enquête statistique s'interprètent aisément comme un processus de récolte, de remontée, de traitement et d'interprétation d'une information diffuse et éparpillée. La réalisation de questionnaires, la formulation des questions posées qui ont vocation à recueillir l'information (plus particulièrement l'utilisation de questions « fermées »), leur codage puis leur saisie (opérations qui permettent de numériser l'information récoltée), le traitement des questions ouvertes et des non réponses et enfin les traitements statistiques – associés ou non à des procédures de redressement – sont autant d'opérations qui, chacune à son niveau, génèrent une quantité certaine de perte d'information. Par ailleurs, les étapes de standardisation des données, d'épurement, de vérifications manuelles et de contrôles de cohérence ont vocation à maîtriser cette déperdition d'information. Mais il reste toujours un degré d'incertitude ou d'erreur, jugé plus ou moins « acceptable », lequel est fonction d'un arbitrage coût/qualité (Rivière [21]). La déperdition d'information issue de ces techniques relève donc de deux natures principales : le recueil des données et leur préparation d'une part, le fondement des traitements sur un processus d'échantillonnage



d'autre part. Ainsi, dans le cadre d'une problématique d'étude par sondages, le phénomène étudié, là encore, n'est pas forcément de nature aléatoire, en revanche l'échantillon l'est par définition.

#### **v) Les plans d'expérience**

Les développements liés aux plans d'expérience s'inscrivent dans une problématique d'expérimentation pour laquelle le praticien ne dispose pas d'un grand nombre d'observations. Soit parce que le support de l'observation est détruit dans le processus, soit parce que cette dernière est très coûteuse à élaborer. L'observation se définit ici comme une réponse expérimentale à la variation d'un ou de plusieurs facteurs. La réponse peut être analysée par une mesure ou différentes procédures de comptage (dénombrement simple, dénombrements ordonnés...). Les secteurs concernés par ces méthodologies relèvent plus particulièrement des domaines agronomique (expériences en champ sur parcelles), industriel (contrôles de qualité) ou médical (expériences sur des traitements en laboratoire).

La planification des expériences consiste à provoquer et à reproduire un phénomène dans le cadre d'un protocole très précis. Il s'agit de répéter et de comparer un certain nombre d'essais ou de manipulations en contrôlant l'influence de facteurs quantitatifs ou qualitatifs, de leurs interactions ou de leurs mélanges éventuels. L'interprétation des résultats se fait à l'aide des modèles d'analyse de la variance à un ou plusieurs facteurs par rapport à une situation en référence. Les tests d'hypothèses d'égalité des moyennes en regard d'une population témoin sont à la base de l'évaluation des résultats des modèles.

Ainsi, l'expérimentation dans ce contexte structuré et contrôlé est une façon de produire de l'observation sur un sous ensemble du réel : celui qui est au centre de l'expérience. Le but est alors de calibrer des effets sur cet espace réduit délimité par les conditions très rigoureuses du dispositif expérimental. Les modélisations mises en œuvre ne se fondent alors pas sur une population d'un grand nombre d'observations individuelles mais sur quelques observations issues des expérimentations. La finalité est bien d'extraire le maximum d'information de la réalisation d'un nombre limité d'expériences. De plus, le choix des observations à recueillir et leur ordonnancement raisonné peuvent se faire dans l'optique de rechercher l'optimalité de la constitution des plans d'expériences. Il s'agit toujours de viser à minimiser la variabilité des prédictions et d'estimer les coefficients des modèles avec le maximum de précision sur l'espace des plans d'expérience retenus.

Dans ce cadre qui utilise des modélisations spécifiques, la perte d'information est imputable à la spécification des variables, de leurs interactions et du modèle par rapport à une distribution jugée aléatoire des données. Par ailleurs, l'élaboration de plans d'expérience optimisés revient à chercher à minimiser un écart au regard de l'influence éventuelle de facteurs incontrôlables sur le phénomène. La possibilité d'extrapoler un résultat final par rapport à des conditions « réelles » beaucoup plus ouvertes et beaucoup plus larges est au centre de la mise en œuvre de ces procédures.

**vi) Les techniques de bootstrap**

Pour lever les interrogations qui quelques fois surviennent à propos de la fiabilité d'une statistique, des techniques particulières, dénommées « bootstrap » peuvent être implémentées. Ces dernières ont pour objet de reproduire un grand nombre de fois la statistique élaborée en faisant varier la composition de la population finie (échantillon) dont elle est issue par ré-échantillonnages successifs. L'idée consiste à mesurer les variations et la sensibilité du résultat. Le but est d'en vérifier sa stabilité (analyse du biais et de l'écart type de l'estimateur) qui peut s'interpréter alors comme une bonne ou une robuste mesure du réel. Dans ce cas, ce qui est mis en évidence, c'est la sensibilité des constructions statistiques à la perte d'information inhérente à l'échantillonnage.

**vii) Les approches dites de « data mining »<sup>1</sup>**

Les progrès accélérés des puissances de calcul des ordinateurs, des capacités de stockage et de la performance des algorithmes ont permis le développement d'outils informatiques regroupés sous le nom de « data mining ». Ces approches adaptées aux grandes bases de données ont pour vocation, littéralement parlant, de fouiller l'univers des données. Elles reposent sur des techniques empruntées à l'analyse exploratoire (exploration de la distribution des variables), aux modèles de régression, aux analyses factorielles, à la classification, aux arbres de décisions ou aux réseaux de neurones.

Ces programmes informatiques ont pour mission de révéler des structures, des associations, des profils ou d'extrapoler et prédire. Il s'agit d'applications de méthodes d'extraction, plus ou moins automatique, d'une information a priori dissimulée sous un amas et un enchevêtrement de données. A la base de ces algorithmes, se trouvent toujours associé l'objectif de maximisation de la restitution d'une information utilisable à travers le prisme de différents indicateurs d'efficacité (tests de significativité paramétrique ou non paramétrique, taux d'erreurs, optimisation de courbes de scores, ...).

Le rôle de filtrage de ces outils est à la fois à motivation algorithmique (gagner du temps de calcul) et exploratoire (éviter de se perdre dans les multiples liaisons et redondances d'information). L'idée pragmatique derrière l'utilisation de ces techniques est que les modèles détectés doivent avant tout coller aux données plutôt que viser l'adéquation à des hypothèses a priori. Dans cette logique, la perte d'information provient de l'écart entre la population exhaustive des données et sa représentation dans une optique d'efficacité opérationnelle (optimisation de plans marketing, ciblage de clientèle, augmentation des ventes croisées ...).

Le caractère automatique et rapide de ces techniques fait qu'elles présentent, vraisemblablement davantage que les techniques plus traditionnelles, le danger de mettre en relief des informations qui peuvent être erronées ou du moins non pertinentes au regard de la problématique visée.

---

1. Cf. numéro spécial du *Journal de la SFdS* (Vol. 142, n°1, année 2001).

La statistique en tant que discipline apparaît par essence réductrice, et sans doute de façon accessoire, d'essence quantitative. La déperdition d'information consubstantielle à la pratique statistique recouvre des natures bien différentes :

- provenant de la distance entre la population mère et la sous population sur laquelle s'appuient les traitements,
- provenant du processus de recueil et de préparation des données,
- provenant des spécifications des traitements (choix d'un modèle, hypothèse de loi, ...),
- provenant de la distribution aléatoire ou incontrôlée des données,
- et évidemment, provenant de la présentation finale des résultats.

Les approches statistiques proposent des mesures de cette perte d'information. Qu'ils se nomment intervalles de confiance, écart type, test d'adéquation, coefficient de corrélation linéaire, pourcentage de variance expliquée, pourcentage de perte d'inertie, coefficients de détermination d'une série temporelle, quantité d'information de Kullback, analyse des résidus d'estimation, ... c'est bien l'objet de ces différents indicateurs que d'apporter au statisticien une estimation de la distance entre un certain « réel » et sa mesure. Le souci inhérent au maniement des statistiques est de toujours maîtriser et minimiser l'éloignement entre les données disponibles au départ et le résultat synthétique final qui est sensé avoir extrait le maximum d'information pertinente eu égard aux objectifs poursuivis (on sait que les critères peuvent être différents selon que l'on se place, par exemple, du point de vue explicatif ou du point de vue prédictif).

Dans le prolongement de ce propos, l'activité du statisticien pourrait alors se caractériser de la sorte : élaborer des méthodologies et des outils de représentation du réel (le plus souvent chiffrée, mais pas nécessairement), en contrôlant – avec le plus de rigueur possible – la perte d'information inévitable. Cette perte d'information restera néanmoins une caractéristique intrinsèque des statistiques puisqu'elle est constitutive du processus de fabrication de l'information. Mais surtout, la déperdition d'information provient du processus intellectuel de simplification indispensable à l'appréhension de la « réalité ».

Les statisticiens, et parmi eux les plus célèbres, se sont depuis longtemps penchés sur ces questions. Sur un plan théorique, par exemple dans le cadre des modèles paramétriques, l'inégalité de Cramer-Rao fournit une borne inférieure – et donc incompressible – de la précision des estimateurs. Sur le plan plus général de la réflexion liée à la connaissance, Fisher (1890-1962) s'est constamment interrogé sur la nature des probabilités, sur l'estimation par rapport à la quantité d'information apportée par l'environnement, sur la notion de population hypothétique infinie; en résumé sur la légitimité de l'inférence par rapport à des constructions assises sur des populations finies. A l'extrême, Karl Pearson (1857-1936) à partir de ses analyses de la corrélation, ira jusqu'à réfuter la réalité profonde des choses et la notion de causalité au profit de l'idée de seules perceptions empiriques construites par le statisticien.

## 5. Les conséquences de cette analyse : réalité ou réalités ?

### 5.1. La statistique assimilée à un langage

Les parties précédentes ont montré que, pour passer de l'infini ou de l'indéfini au fini, pour construire un ordre nécessaire pour mener à bien sa pensée, il y a nécessité de procéder à des partitions du réel. Mais peut-être le réel n'existe pas. Seules des perceptions de la réalité existent, et leur élaboration nécessite de passer par des découpages ou des catégorisations. A l'image de la production du peintre, le processus de découpage s'accompagne d'un processus de recomposition. Ces opérations de la pensée ne sont pas sans impact sur la perception des objets que l'on construit. Elles induisent une subjectivité consubstantielle à la perte d'information qui découle de la recherche de résumés : il ne sera jamais possible de tout classer de manière précise, sauf à tomber dans l'énumération exhaustive.

Selon la thèse de Michel Foucault, une similitude se dégage entre la fonction essentielle du langage et ce mode d'approche du réel intrinsèque aux sciences. Ainsi, l'histoire des sciences naturelles (pris dans ce cadre de réflexion à titre d'illustration), et peut-être pourrait-on ajouter, le mécanisme de catégorisation ou de classification du réel, « trouve son lieu dans cette distance maintenant ouverte entre les choses et les mots – distance silencieuse, pure de toute sédimentation verbale et pourtant articulée selon les éléments de la représentation, ceux-là même qui pourront de plein droit être nommés. Les choses abordent jusqu'aux rives du discours parce qu'elles apparaissent au creux de la représentation. »

Le diagramme de la figure 1 basé sur un vocabulaire emprunté à l'analyse sémiologique, propose de visualiser les différentes sources de subjectivité ; celles inhérentes à la perte d'information (passage de l'objet au concept via le signe) et celles associées à la recherche du sens (passage du concept à l'interprétation).

Dés lors qu'intervient une information chiffrée (une statistique), le processus de pensée met en jeu ces différentes sphères. Entre chacune se nichent des espaces potentiels de subjectivité. Aussi, par analogie, on peut considérer que l'activité statistique s'assimile à un langage, et comme tout langage, ne peut échapper à la subjectivité.

### 5.2. Quelques sources de subjectivité liées au processus de catégorisation du réel

Aucun découpage n'a de validité universelle. Le schéma suivant relève que les sources de subjectivité interviennent en différentes phases du processus de pensée. Elles peuvent être de natures bien différentes. Outre le passage au signe – dans notre domaine le seul fait de produire une statistique – d'autres aspects sources de subjectivité méritent d'être mentionnés.

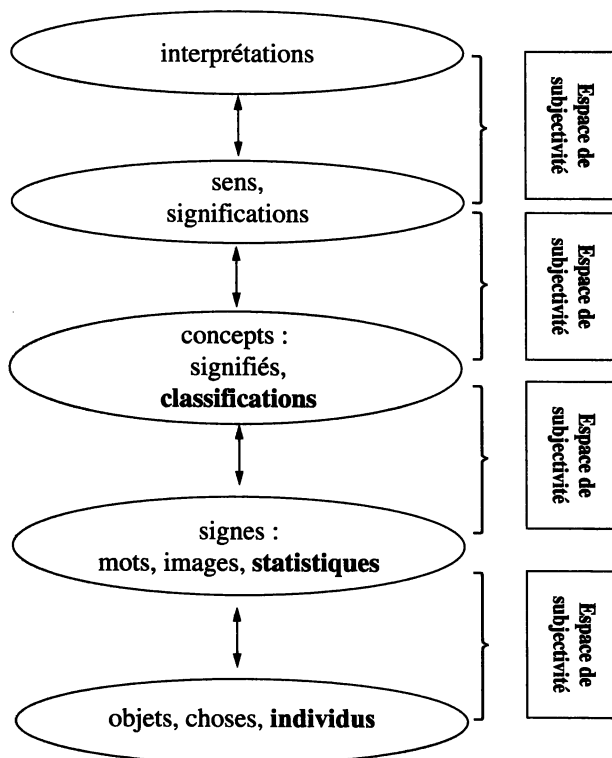


FIG 1. — Des objets à leurs interprétations.

*Guide de lecture du graphique :*  
 Le chemin de l'appréhension de l'objet à son interprétation laisse la place à de multiples espaces de subjectivité. En effet, pour passer des choses ou des objets à leurs interprétations, tout langage utilise des signes (que sont les mots, les images, ou pour ce qui nous concerne ici, les statistiques), lesquels donnent une texture aux concepts (les signifiés) lesquels à leur tour, permettent de donner du sens et d'ouvrir à l'interprétation.

*i) La présentation d'un résultat*

Une fois la palette d'indicateurs statistiques constituée, il faut choisir les résultats chiffrés à restituer et la façon de les présenter. Cela peut être sous forme rédactionnelle, de tableaux, de graphiques, de cartes, d'annuaires, de fichiers. Quel que soit le mode de présentation, celui-ci induit une perte d'information par rapport à la production totale de statistiques élaborée au cours de la phase étude. De même, les choix de présentation déterminent un angle de vue qui filtre et oriente l'analyse par les éléments qu'il met en relief.

*ii) L'environnement culturel*

Une source de subjectivité est éminemment liée à l'environnement culturel. Georges Perec nous rappelle que, pour désigner l'espace qui sépare les constructions dans nos villes, il y a environ sept termes en français (rue, avenue, boulevard, place, cours, impasse, venelle) quand il y en a au moins vingt en anglais (street, avenue, crescent, place, road, row, lane, mews, gardens, terrace, yard, square, circus, grove, court, greens, houses, gate, ground, way, drive, walk).

*iii) La relativité temporelle*

L'évolution de l'état des connaissances impacte la pertinence d'une classification comme le montre l'histoire de la taxinomie en biologie. Les classifications des naturalistes se perfectionnent sans cesse, sous l'impulsion de découvertes d'espèces d'animaux ou de plantes hier inconnues, du développement des hybrides, de l'émergence de nouvelles technologies ou de l'évolution des méthodes taxinomiques.

Les statisticiens ont toujours été sensibles à ces questions. À titre d'exemple Desrosières, dans un ouvrage très documenté [7], analyse et met en perspective historique l'évolution de nomenclatures économiques et son lien avec les transformations dans les domaines de l'administration, de l'action publique ou de l'ordre social observées notamment en France et en Grande-Bretagne au cours des XVIII<sup>e</sup>, XIX<sup>e</sup> et XX<sup>e</sup> siècles.

*iv) La finalité visée*

La subjectivité peut résulter des objectifs assignés à une classification. Doit-on privilégier l'aspect pratique : ranger l'information pour permettre d'aller chercher rapidement les éléments, ou doit-on chercher avant tout à donner du sens aux concepts pour favoriser la réflexion (et négliger indirectement, dans certains cas, les aspects pratiques concernant l'accès aisé aux éléments unitaires) ?

Compte tenu des lois de la combinatoire le nombre de relations d'ordre étant très élevé, toute tentative de classement à vocation totale serait vaine. C'est cet aspect que déplore Georges Perec dans son texte « Penser/Classer ». Ainsi en est-il de ce que ressent le lecteur qui ne réussit pas à se déterminer sur un thème de classement pour sa bibliothèque personnelle : l'ordre alphabétique de l'auteur, le titre de l'ouvrage, le genre, la taille ? Quel que soit le mode de classement retenu celui-ci n'est jamais total, il génère toujours un livre ou un type d'ouvrage inclassable. Dans un rangement par ordre alphabétique de l'auteur, comment alors classer les livres de cuisine nous demande Perec ? Comment classer les ouvrages collectifs ?

*v) Le traitement de l'absence d'information*

Un ensemble peut se définir par les éléments qui le composent, mais aussi par ce qu'il n'est pas. Ce qui n'entre pas dans une catégorie peut être tout aussi important que ce qui est catégorisé. En termes d'exploitations statistiques, les

catégories « autres », « divers », « ne sait pas », ou « non renseignées » jouent un rôle qui n'est pas neutre dans nos problématiques et qui, par voie de conséquence, sont souvent sujettes à d'âpres discussions.

*vi) L'imbrication du sujet dans la perception de l'objet : l'effet retour*

La problématique est historiquement identifiée par les statisticiens. Ainsi Cournot [15], dans ses écrits liés à la statistique (1843), considère que l'implication du sujet dans le processus de perception et de mesure de l'objet passe par la notion de « probabilités philosophiques ». Ces dernières, en quelque sorte, constituent le socle, l'outillage a priori du statisticien. Celui-ci s'efforcera d'objectiver les présupposés ou attendus par la mesure statistique. La mesure de l'objet, du phénomène dépend néanmoins toujours des hypothèses sous-jacentes liées à ces « probabilités philosophiques » qui sont fonction des connaissances initiales.

Toute activité de catégorisation oriente le processus de pensée et induit donc un effet retour sur la perception du monde qu'elle génère. Ce n'est pas seulement l'objet qui est appréhendé par l'esprit. Le processus inhérent à la pensée oriente la perception que l'on se forge de l'objet. Un classement alphabétique est souvent associé à l'idée de hiérarchisation. La classe des vertébrés mammifères marins (baleines) revient à accorder une prédominance au critère de mode de reproduction des animaux plutôt, par exemple, qu'à leur milieu d'évolution naturelle (qui les rapprocherait plus des poissons). L'usage de statistiques génère cet effet retour sur l'objet pensé. Ce point, essentiel dans le domaine de la statistique publique, est rappelé par Padiou [18] qui souligne à cet égard la responsabilité du statisticien dont la production « structure les représentations sociales ».

*vii) La frontière avec l'imagination*

Par ailleurs, le mécanisme de classification étant intimement imbriqué dans celui de la recomposition, toute représentation d'une réalité effleure l'imaginaire. La frontière entre un réel recomposé et l'imaginaire peut être ténue, à l'image des chimères<sup>2</sup> de la mythologie (le centaure, le griffon, le sphinx). Aussi, une classification peut créer des imprécisions, des incohérences ou des contradictions jusqu'à éventuellement générer du non-sens : telle la fameuse encyclopédie chinoise de Borges, qui mêle entre autres, animaux mythiques et animaux réels.

*viii) Le passage du tout à la partie, de la partie au tout*

Le processus de catégorisation peut aller, comme le montre Boris Vian avec dérision, jusqu'à créer une vision artificielle de la réalité. C'est ce qu'évoque la situation absurde relatée dans *l'Automne à Pékin*. Un des personnages du roman, l'archéologue, réalise des fouilles dans un désert. A chaque découverte de vase antique il s'empresse de le briser en de multiples morceaux à l'aide de son marteau afin de pouvoir remplir, au mieux, tous les compartiments

---

2. La montée vers l'imaginaire peut se faire par paliers, l'animal chimérique pouvant l'être à un deuxième degré, à l'instar de l'hippogriffe, croisement du cheval avec le griffon.

de ses boîtes standards d'archéologie. Une telle situation relève bien d'une caricature extrême puisqu'en l'occurrence le découpage du réel aboutit à produire un niveau d'information bien inférieur à ce qu'une vision globale de l'objet aurait pu rendre. Le caractère unique, si précieux, de l'objet disparaît pour les « besoins » de l'approche scientifique.

Cette parabole rejoint ici la critique d'Edgar Morin considérant que la statistique concourt à un mode de pensée qui soit « unifie abstraitement en annulant la diversité », soit « juxtapose la diversité sans concevoir l'unité ». Elle a ainsi tendance, selon sa vision de la complexité du réel, à générer une intelligence aveugle, une pensée réductionniste et simplifiante qui « ignore la réalité du système abstrait d'où sont les éléments à considérer »<sup>3</sup>.

Là encore, les statisticiens se sont depuis longtemps interrogés sur le sujet. La notion « d'homme moyen » de Quetelet (1796-1874) et les débats sur la « physique sociale » auxquels il participait, sont une illustration des tensions permanentes entre les approches de l'individuel et de l'universel, qui ont accompagné le développement des outils statistiques [7].

Le processus intellectuel de catégorisation, à certains égards, s'oppose à une approche des choses et des êtres, plus globale ou plus liée à l'individualité, qui relèverait plutôt des sentiments ou de la poésie. Sous une autre forme, ce dialogue du Petit Prince de Saint-Exupéry renchérit dans cette idée : « Les grandes personnes aiment les chiffres. Quand vous leur parlez d'un nouvel ami elles ne vous questionnent jamais sur l'essentiel. Elles ne vous disent jamais : « Quel est le son de sa voix ? Quels sont les jeux qu'il préfère ? Est-ce qu'il collectionne les papillons ? » Elles vous demandent : « Quel âge a-t-il ? Combien a-t-il de frères ? Combien pèse-t-il ? Combien gagne son père ? ». Alors seulement elles croient le connaître. »

## 6. Éléments de conclusion

Ces réflexions amènent à quelques considérations sur l'idée que l'on peut se faire du métier de statisticien.

Les travaux statistiques se formalisent via un langage mathématique qui sert de support à la construction et à l'utilisation des concepts. Quel que soit le niveau de complexité des algorithmes et processus utilisés, pour autant, il ne faut pas que cette terminologie scientifique masque la part de subjectivité associée à notre discipline. Fisher [10] a souvent souligné cette différence d'approche entre le mathématicien (qui délivre des résultats exacts) et le statisticien qui doit inférer dans un contexte d'incertitude. Le véhicule mathématique des théories et pratiques statistiques, comme tout langage scientifique, a pour finalité de réduire le champ des interprétations possibles pour tendre vers une « objectivité ». Il est assurément un gage de rigueur et permet de limiter l'importance de la subjectivité, il ne la supprimera

---

3. Mais cette vision ne consiste qu'à mettre en évidence sur un plan critique ce qui relève d'une dimension essentielle de la discipline statistique : proposer des constructions qui permettent de « faire de l'un à partir du multiple » et de « défaire cette unité pour refaire de la diversité » (Desrosières. A [7] p. 87).





pourtant jamais mais il permet de mieux la partager. Le domaine médical est une bonne illustration de cette fonction de «subjectivité partagée» induite par le langage statistique. En médecine, l'utilisation de statistiques, via les processus de codage et de détermination de nomenclatures, assure le passage de l'analyse individuelle (à des fins de diagnostic) à l'analyse catégorielle (dans une perspective d'études épidémiologiques et de la mise en place d'actions de prévention par exemple). Dans le cadre de cette opposition de vues, la construction de statistiques médicales a été à l'origine de nombreux débats entre médecins statisticiens, notamment au XIX<sup>e</sup> siècle [7]. Toujours est-il que l'utilisation de nomenclatures a créé un langage commun qui permet d'asseoir un dialogue entre les praticiens, de transmettre les connaissances et de mesurer l'efficacité des thérapeutiques.

Au regard d'une réalité quasi infinie, la perte d'information induite par l'utilisation des techniques statistiques est donc consubstantielle à notre discipline. Pour extraire et mettre en relief des informations pertinentes, il faut accepter d'en perdre sur quelques dimensions périphériques. Simplement, le statisticien doit gérer au mieux, avec la plus grande rigueur intellectuelle et morale, cette déperdition d'information. Si les outils statistiques nous apportent une meilleure lisibilité du réel, compte tenu néanmoins de leur nature subjectivè, ils génèrent un espace qui laisse une place importante aux questions de déontologie. Quelle que soit la source de la subjectivité, elle engage la responsabilité individuelle du statisticien dans la mesure où celui-ci se retrouve souvent seul face à ses choix. C'est la raison pour laquelle il y a utilité au travers de la rédaction de chartes ou de codes, de guider le statisticien en édictant «des principes auxquels l'ensemble de la profession adhère» (Padieu [18]).

Deux types principaux d'orientation déontologique, aujourd'hui admis par la profession peuvent être mentionnés. Ils consistent :

- à considérer que les statistiques produites doivent être «dignes de foi, objectives», «les statistiques doivent refléter la réalité de façon exacte et fiable» (cf. principe 12 du Code des bonnes pratiques de la statistique européenne [5]).
- à considérer que «toute mesure statistique, toute interprétation et analyse comporte une marge d'incertitude et implique un choix entre points de vue.» (cf. alinéa 6 du Code de déontologie de l'association des administrateurs de l'INSEE [6]).

À l'aune des réflexions précédemment évoquées, la dichotomie proposée par Desrosières ([7], [8]) met en avant deux approches épistémologiques qualifiées respectivement de «réaliste» et de «conventionnaliste» et montre que le statisticien navigue toujours d'une berge à l'autre. L'analyse menée supra conclurait plutôt que l'essence de la statistique verse vers le conventionnel du langage (utilisation de signes communément partagés). Le seul statut épistémologique alors tenable est de considérer que la statistique ne peut jamais se prévaloir d'exprimer le **réel** mais plutôt d'exprimer ou de mesurer **une réalité**. La démarche menée ici opte donc pour une épistémologie radicalement «conventionnaliste».

Aussi la déontologie par rapport à la recherche de qualité peut relever plus du « partage des signes » et donc de la rigueur sous-jacente, de la transparence, des efforts de présentation, d'explication et de vulgarisation que de déclarations d'intentions sur le caractère « objectif » d'une production de statistiques.

Finalement, on pourrait paraphraser Roland Barthes qui disait à propos du langage de l'écrivain, qu'il « ... n'a pas à charge de *représenter* le réel, mais de le signifier », en étendant cette citation au langage du statisticien. Au reste, c'est bien ce principe que le Code des bonnes pratiques de la statistique européenne entend mettre en avant en demandant le respect du critère de pertinence dévolue à la production d'une statistique : « les statistiques disponibles doivent correspondre aux besoins des utilisateurs ». Il s'agit avant tout d'affirmer que la fonction la plus fondamentale de la production statistique est de donner du sens.

Produire des statistiques apparaît alors indissociable, par essence, d'un engagement éthique, lequel provient de cette distance incompressible qu'il y aura toujours entre le « réel », sa représentation et son interprétation. De même que chez Descartes la réflexion sur la finalité des sciences s'accompagnait d'interrogations relatives au doute, l'activité de statisticien ne doit jamais s'éloigner, à mon avis, de ces préoccupations.

## Bibliographie

- [1] BARTHES R. (1957). *Mythologies*. Le Seuil, Paris.
- [2] BORGES J.-L. (1944). *Fictions*. «Bibliothèque de la Pléiade». Gallimard, Paris 1993.
- [3] BORGES J.-L. (1952). *Autres Inquisitions*. «Bibliothèque de la Pléiade». Gallimard, Paris, 1993.
- [4] BORGES J.-L., GUERRERO M. (1969). *Le livre des êtres imaginaires*. Gallimard, Paris, 1987.
- [5] *Code de bonnes pratiques de la statistique européenne* (2005). Reproduit dans le Journal de la Société Française de Statistique, 147, 2,
- [6] *Code de déontologie statistique* – proposé par l'association des administrateurs de l'INSEE, (avril 1984).
- [7] DESROSIÈRES A. (2000). *La politique des grands nombres. Histoire de la raison statistique*. La Découverte/Poche, Paris.
- [8] DESROSIÈRES A. (2003). La qualité des quantités. *Courier des Statistiques*, n° 105-106, 51-62
- [9] DESCARTES R. (1637). *Discours de la méthode*. Librairie Générale Française, Paris.
- [10] FISHER R.A. (1936). Uncertain Inference. *Proceedings of the American Academy of Arts and Science*, 71, 245-258.
- [11] FISHER R.A. (1958). The Nature of Probability. *Centennial Review*, 2, 261-274.
- [12] FOUCAULT M. (1966). *Les mots et les choses. Une archéologie des sciences humaines*. Gallimard, Paris.

- [13] GOURIÉROUX C. et MONFORT A. (1989). *Statistiques et modèles économétriques Vol.1 – Notions générales, Estimations, Prévision, Algorithmes*. Economica, Paris.
- [14] JERPHAGNON L. (1989). *Histoire de la pensée : Antiquité et Moyen Age*. Tallandier, Paris.
- [15] MARTIN T. (2004). *Cournot. Statistique et raison des choses in Histoires de probabilités et de statistiques*, coordonné par Barbin E. et Lamarche J-P. Ellipses, Paris.
- [16] MAYR E. (1982). *Histoire de la biologie : diversité, évolution et hérédité*. Fayard, Paris.
- [17] MORIN E. (1990). *Introduction à la pensée complexe*. Le Seuil, Paris.
- [18] PADIEU R. (1999). La déontologie statistique. *Journal de la Société Française de Statistique*, 140, 1, 5-21.
- [19] PEARSON K. (1912). *La Grammaire de la Science*. Alcan, Paris.
- [20] PEREC G. (1985). *Penser/Classer*. Le Seuil, Paris.
- [21] RIVIÈRE P. (2003). Approche coût-qualité pour l'amélioration des processus de production statistique. *Courrier des statistiques* n° 105, 65-75.
- [22] SAINT-EXUPÉRY A. (1946). *Le Petit Prince*. Gallimard, Paris.
- [23] VIAN B. (1956). *L'Automne à Pékin*. Les éditions de minuit, Paris.
- [24] ZERNER M. (2004). *Statistique et modèles probabilistes de Fisher à Havelmoo in Histoires de probabilités et de statistiques*, coordonné par Barbin E. et Lamarche J-P., Ellipses, Paris.