Le 22 septembre 2003

Quelques réflexions sur la « Star Academy » dans le cadre du RIS :
(D. Deprins)

§1 Le choix de l'objet « Star Academy » : un léger malaise …

Force est de constater que la « Star Academy » est un sujet qui fait couler beaucoup d'encre ou, à tout le moins, fait énormément parler de lui.

J'avoue ressentir un léger malaise par rapport à ce choix ! Et c'est ce malaise qui m'a fait repenser au récent livre d'E.-E. Schmitt[1] : « Lorsque j'étais une œuvre d'art ». En deux mots, il y est question d'un jeune homme de vingt ans au bord du suicide, tout à sa folie d'être tout par lui-même. Un artiste en quête de scandale lui offre un marché de dupe : transformer son corps en œuvre d'art, en objet d'admiration… Et le jeune homme d'accepter ! On imagine sans peine toutes les dérives possibles : en tous les cas, la privation et la perte de liberté qui résultent de son instrumentalisation déshumanisante. Et le public d'en être complice !

Et c'est là que ce livre (qui pousse loin, très loin la caricature !) m'a amenée à ce spectacle de télé-réalité qu'est la « Star Academy ». Je ne peux m'empêcher de m'interroger sur l'éventuelle caution que nous apporterions, nous téléspectateurs, à l'instrumentalisation de ces jeunes participants. Ce n'est pas un scoop de dénoncer que les producteurs et organisateurs en tous genres se moquent bien de ce qu'ils pensent et ressentent ; plus encore, ils utilisent les émotions de ces « stars » de quelques saisons (gros plan sur les larmes, sur les manifestations de découragement, sur les mouvements d'humeur et de jalousie etc.) pour asseoir la fidélité, voire l'addiction, des téléspectateurs. Et surtout, tout ce que cela impliquerait de leur comportement de consommateurs (journaux, CD, émissions et autres objets corrélés…). Ils se jouent et jouent ainsi de leur *imaginaire*[2].

Instrumentalisé aussi, le téléspectateur… Pas tous, bien sûr … mais voici une petite anecdote qui laisse songeur ! Lors d'une information express à la radio, un commentateur reparlait de la lauréate *Jennifer* de la première « Star Academy » et vantait le succès de ses tournées de chants dans les grandes villes francophones de notre pays. Il s'étonnait de sa voix grave de rockeuse qui contrastait quelque peu avec l'âge si tendre de son public âgé de … 5 à 14 ans qui fait salle comble lors des concerts de son idole ! Rassurant ou inquiétant ?

[1] SCHMITT E.-E., *Lorsque j'étais une oeuvre d'art* (2002), Albin Michel.
[2] Cet imaginaire est à comprendre dans le sens très particulier de son acception lacanienne. L'**imaginaire** est un terme fondamental de la terminologie et de la théorie lacanienne qui forme avec les deux autres registres du symbolique et du réel, une triade indissociable.

On pourrait en dernier ressort espérer que cette émission fasse naître sous nos yeux de jeunes artistes, ne fût-ce que l'un ou l'autre, de temps en temps… Mais être Star, avec son aspect éphémère, ce n'est pas être artiste. Un artiste crée et, le plus souvent, le temps y participe ; il sort du sillon du « déjà vu », du « déjà entendu ». Quelque chose du réel qui nous avait échappé ou qu'on ne connaissait pas est dévoilé, comme la carnation de la peau est révélée par les peintres du XVII^{è} siècle. Ici, qu'y a t-il de l'ordre de la création ? Des chansons entendues mille fois? Pas de voix vraiment originales, ni d'interprétations magistrales. Ou en tous les cas, pas de place laissée à de telles émergences ; émergences qui semblent désespérément cadenassées par un corps professoral caricatural, abrasant le désir de création de l'élève pris dans ce rapport de fascination.

Et qui plus est, le temps qui permettrait peut-être une telle émergence n'y est pas : il faut des résultats visibles et même, rapidement visibles !

En plus, c'est la célébration du laid, du mauvais goût, des paillettes et du faux ; célébration à la Berlusconi, sorte de jongleur entre pouvoirs économique, politique et médiatique. A. Tabucchi[3], auteur contemporain italien, écrit que « …l'esthétique[4] est directement liée à l'éthique et à la politique, l'esprit de certaines idées politiques se manifestant toujours à travers des choix esthétiques. Malheureusement, par moment, cette deuxième composante domine, ainsi depuis quelques temps en Italie, on assiste à la célébration du laid … ». Et Berlusconi de donner une représentation politique à cette tendance. Et en Belgique ?

§2 Le petit bout de la lorgnette du statisticien-probabiliste :

Pour aborder la point de vue du statisticien-probabiliste, il fallait faire un choix : c'est celui de l'articulation de cette discipline qui fut retenue parce qu'il offre la possibilité d'en faire comprendre les raisonnements, les concepts de base, ses possibles et ses limites. Cette articulation qui noue, diversement d'ailleurs, la Statistique Descriptive, la Théorie des Probabilités et l'Inférence Statistique sera d'abord envisagée de façon tout à fait générale pour être ensuite appliquée à notre objet, la Star Academy.

§ 2.1 *De l'articulation théorique entre la Statistique Descriptive, la Théorie des Probabilités et l'Inférence Statistique* :

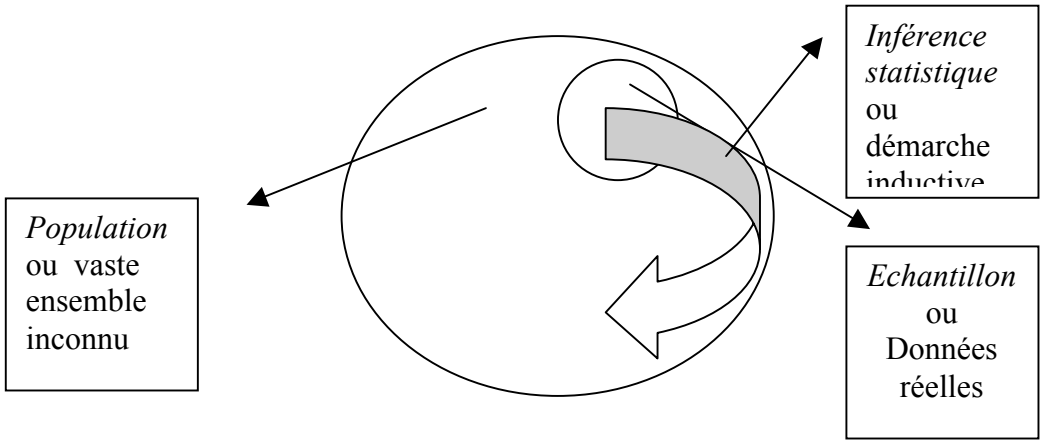
Le mot « statistique » a, en français moderne, au moins deux sens distincts. Pour le public et pour la presse, il désigne des mesures portant sur le monde économique et social, telles que le taux de chômage, l'indice des prix, les taux de mortalité, les intentions de vote, etc. Ce sont principalement « les » statistiques des Instituts de Sondages ou des Instituts Nationaux de Statistiques . Mais dans une université ou dans un laboratoire de recherche scientifique, il correspond à une branche des mathématiques appliquées, apte à modéliser des phénomènes où interviennent de grands nombres de mesures, qu'elles soient sociales (Sondages Politiques, Démographie, Psychologie, Criminologie,…), économiques (Finance, Marketing, Analyse Economique, …), médicales (Analyse des Données de Survie, Epidémiologie,…), relatives à l'énergie (Consommation d'Energie, Comparaison des Sources

[3] TABUCCHI A. *Lire,* Interview juin 2002.

[4] *Le petit Larousse 2002* **:** « L'esthétique *»* : 1. Théorie du beau, de la beauté en général et du sentiment qu'elle fait naître en nous. 2. Ensemble des principes à la base d'une expression artistique, littéraire, etc… visant à la rendre conforme à un **idéal** de beauté.

Energétiques,…), aux sciences naturelles (Ecologie, Ressources Géologiques,…), à l'ingénierie (Contrôles de Qualités, …), etc. C'est « la » Statistique ou Statistique Mathématique ou encore l'Analyse Statistique.

L'Analyse Statistique est donc un ensemble de méthodes mathématiques qui , à partir du recueil et de l'analyse de données réelles, permettent l'élaboration de modèles probabilistes autorisant des estimations et des prévisions. C'est ce que tente d'illustrer, de façon très schématique, la figure ci-dessous.



Cette définition, pour le moins compacte, dénonce d'emblée la démarche inductive de l'Analyse Statistique appelée, dans son jargon, *l'inférence*[5] *statistique*. Dans le courant « expérimentalo-inductif » de l'épistémologie des sciences, cette démarche inférentielle consiste à tirer des conclusions à partir de ces données réelles observées (ou *échantillon)* sur un plus vaste ensemble inconnu (ou *population*)- dont les données réelles dont on dispose ne sont qu'une information limitée. Passer ainsi du particulier au général introduit par nature de l'incertitude, du biais, du hasard dans son sens épistémologique d'aléatoire.

L'incertitude, le hasard, l'aléatoire sont donc au centre des préoccupations du statisticien-probabiliste ; une des façons de lire l'histoire de la statistique[6] est celle d'une « domestication du hasard », selon la belle formule de Hacking[7] . Stigler associe la statistique à la mesure de l'incertitude, comme le dit le titre de son livre[8]. Et toute la construction méthodologique de l'Analyse Statistique tentera de prendre en compte et de rendre compte de ces éléments – incertitude et hasard –, sans jamais les éradiquer.

---

[5] *L'inférence* est à comprendre ici dans le sens plus restrictif de l'induction. En logique, la déduction est aussi une inférence.

[6] DESROSIERES A « La diversité et l'enchevêtrement des origines de la statistique expliquent que son histoire puisse être racontée de diverses façons : sociologique, institutionnelle, mathématique et philosophique » dans le *Dictionnaire d'Histoire et de Philosophie des Sciences* de D. Lecourt, PUF, Paris, 1999.

[7] HACKING I.,*The taming of chance*, Cambridge, cambridge University Press, 1990.

[8] STIGLER S., *The History of Statistics. The Measurement of Uncertainty Before 1900*, Cambridge (Mass), Harvard University Press, 1986.

Dans son chapitre sur « Le hasard désacralisé : de l'oracle du tirage au sort à l'indifférence de l'aléatoire », H. Atlan[9] rappelle que « le hasard n'a pas toujours été synonyme d'absence de sens, d'aléatoire au sens d'indifférent . Bien au contraire, il joue souvent dans la pensée magique le rôle d'oracle par lequel le dieu se manifeste. Le tirage au sort permet de combler l'ignorance et de décider en situation d'incertitude ».

Cependant, l'usage de la notion de hasard est presque toujours lié à une difficulté de l'explication causale. Bien que la notion de hasard soit quelque peu ambiguë et vague, on peut tout de même tenter de lui donner une signification méthodologique précise. Dans le *Dictionnaire d'Histoire et de Philosophie des Sciences*[10], J Gayon propose trois sens épistémologiquement pertinents du mot « hasard » : la chance, la contingence par rapport à un système théorique et l'aléatoire. Cependant si l'on s'en tenait à son origine étymologique (az-zahr, « le dé » en arabe), cette étymologie nous conduirait à restreindre le hasard à l'une de ses significations, l'aléatoire (du latin *alea*, « dé »), récusant du même coup le « hasard » comme terme générique.

Des trois sens épistémologiquement pertinents du mot « hasard » - la chance, la contingence par rapport à un système théorique[11] et l'aléatoire - seul celui de l'aléatoire est convoqué par l'Analyse Statistique et la Théorie des Probabilités que la première convoque à son tour.

Rappelons que la notion de hasard comme « chance » est son sens le plus familier et le plus populaire ; il apparaît dès lors que quelque chose se produit alors qu'on ne l'attendait pas et qu'aucun plan d'action défini à l'avance n'en ait « programmé » l'occurrence. Aujourd'hui, on dirait : « il (ou elle) a eu de la chance ». C'est la *tuchè* (la fortune) et *to automaton* (le hasard) d'Aristote [12] lorsqu'une fin est atteinte sans avoir été la cause immanente de l'effet produit. Le deuxième sens épistémologique du mot « hasard » est celui de la « contingence par rapport à un système théorique ». Cette acception du mot « hasard » concerne surtout la physique et les sciences biologiques, ce qui nous éloigne de notre propos.
Enfin, l' « aléatoire » de l'*Analyse Statistique* (et du *Calcul des Probabilités*) est le troisième sens, plus technique, et beaucoup plus tardif, du mot « hasard ». Comme l'explique Nagel [13], il s'applique à des événements dont on ignore absolument les conditions déterminantes (ce qui ne signifie pas qu'il n'y ait pas de conditions déterminantes) ou, plus exactement, des événements dont on sait qu'ils peuvent se réaliser dans certaines classes de conditions, ou catégories, mais tout en ignorant lesquelles sont effectivement réalisées dans un cas particulier. Le jeu de pile ou face est l'archétype d'une situation aléatoire. Une pièce de monnaie qui tombe sur « pile » ou sur « face » est un événement aléatoire ; la survenue

---

9 ATLAN H., *Les étincelles de Hasard*. tome 1 : *Connaissances spermatiques*, Ed du Seuil, Librairie du XX^e^ siècle, 1999.

[10] LECOURT D., *Dictionnaire d'Histoire et de Philosophie des Sciences,* PUF, Paris, 1999.

[11] Comme le souligne J. Gayon, le terme « fortuit » que l'on confond souvent avec les trois sens épistémologiques donnés du mot « hasard », serait d'avantage à considérer comme un terme générique que comme un sens particulier du mot « hasard », même si son étymologie (*fors,* la « fortune » en latin) fait penser à l'un des sens du mot – la « chance », p 475-476.

[12] Bien entendu, il y a une distinction entre la *tuchè* et l'*automaton*, distinction aristotélicienne qui relève essentiellement du domaine d'application de la notion. La *tuchè,* limitée à la *praxis,* est un cas particulier de *to automaton.*

[13] NAGEL E., *The Structure of Science. Problems in the Logic of Scientific Explanation*, Londres, Routledge & Kegan, 1961.

singulière d'un des deux résultats possibles – pile ou face –ne peut être prédite avec exactitude.
Cette notion du hasard requiert d'avantage de théorie que la première en ce sens qu'il faut qu'une hypothèse soit faite sur ce qui est aléatoire, justement. C'est ici qu'intervient la *Théorie des probabilités*[14] (ou *Calcul des Probabilités*) qui suppose que les événements se produisent dans certaines conditions, avec une certaine fréquence. Un événement aléatoire est un événement qui suit une « loi de probabilité », liant ainsi inexorablement l'aléatoire à une théorie formelle qui lui donne son sens .

« L'aléatoire moderne est le plus souvent considéré comme ce qui se produit *sans cause* connue, ni même connaissable et dont la survenue singulière ne peut donc être prédite[15] ». Plus rigoureusement, cette notion de hasard n'exclut pas qu'il y ait *de facto* une causalité stricte au niveau de l'événement : il y a sûrement tout un jeu de causes strictement déterministes qui aboutit au résultat « pile » du jet d'une pièce équilibrée plutôt qu'au résultat « face ». Le mouvement de la pièce, corps soumis au champ de la pesanteur et astreint à évoluer au-dessus d'une surface plane comme une table, devrait en principe être entièrement calculable à partir des lois de la mécanique. *Mais ce calcul est en pratique impossible à mettre en œuvre, pour la raison que des modifications imperceptibles dans les conditions du mouvement (manière de lancer, point de chute, ...) peuvent en changer radicalement l'aboutissement final*[16] De façon générale, faute de théories et/ou d'observations suffisamment précises, on ne connaît pas ces causes. Ainsi la chute d'une pièce équilibrée sur le côté « pile » est-elle rigoureusement déterministe[17]. Dans un souci d'efficacité, mieux vaut renoncer à une prédiction par les lois physiques : la seule prédiction possible pour de tels événements reste alors la prédiction statistique qui raisonne directement sur les caractéristiques du hasard lui-même. La statistique et les probabilités constituent alors une méthode puissante pour tenter de mesurer le hasard afin de le maîtriser, de le « domestiquer ».

Pour pouvoir tirer des conclusions inférentielles sur l'ensemble inconnu – la *population* qui est l'objet de son intérêt – à partir des données réelles de l'*échantillon* dont il dispose, le statisticien aura d'abord à préparer, analyser et apprendre à connaître ces données réelles qui constituent sa matière première. C'est le passage obligé de la « *Statistique Descriptive* » qui n'a d'autre prétention que de « décrire », comme son nom l'indique, ce matériel de base du statisticien. La *Statistique Descriptive* offre une panoplie d'outils utiles à la description des données réelles. C'est aussi ce qu'on appelle l'*Analyse des Données* ou l'*Analyse Exploratoire des Données* dans sa version multidimensionnelle des grands tableaux de données collectées. Des simples indicateurs de tendance centrale, comme la moyenne arithmétique, ou de dispersion, comme l'écart-type, aux Analyses Factorielles des Correspondances Multiples, en passant par les Analyses Discriminantes, Analyses de Classification, de Corrélations ou les Histogrammes, Stéréogrammes et « Box-Plot » etc., cette première analyse incontournable consistera toujours à *décrire, résumer et visualiser l'information* recelée dans les données réelles disponibles. Y décèlera-t-on une structure cachée des données ou des valeurs particulières « aberrantes » qu'il faudra supprimer ?

---

[14] La *Théorie des probabilités* (ou *Calcul des Probabilités*) a pris aujourd'hui son indépendance complète et peut être enseignée comme une branche des mathématiques sans référence au concret ; elle reste une formalisation du réel, avec comme toute formalisation, l'inconvénient de la simplification et de la distorsion mais aussi l'avantage de permettre à l'esprit humain d'appréhender et mieux agir sur le réel.
[15] ATLAN H., *Les étincelles de hasard*, Op. Cit., p.351.
[16] DIU B. et alii, *Elements de Physique Statistique*, Paris, Hermann, 1989, appendice I.
[17] Notons toutefois que ce sens du mot « hasard » comme événement aléatoire n'implique pas nécessairement qu'une telle causalité existe (exemple de la mécanique quantique).

Afin d'aller au-delà de la simple description des données de la Statistique Descriptive, la *Théorie des Probabilités* (ou le *Calcul des Probabilités)* permettra de faire une hypothèse « stochastique » d'un modèle probabiliste sur la façon dont les données réelles observées auraient été générées : en d'autres termes, elle propose des lois de probabilité, modèles abstraits, qui permettrons d'associer à l'occurrence d'un événement aléatoire, une mesure de sa probabilité - exactement comme on associe au résultat « face » d'une pièce de monnaie équilibrée, une probabilité de ½ (comme pour le résultat « pile », évidemment !). Cette démarche s'inscrit cette fois dans l'autre courant de l'épistémologie des sciences, celui « hypothético-déductif ». C'est d'ailleurs cette tension interne entre ces deux courants épistémologiques « hypothético-déductif » et « empirico-inductif » qui fait la richesse des interprétations possibles du raisonnement probabiliste et statistique. La *Théorie des Probabilités* fournit donc l'outil mathématique, et abstrait, qui permet de mesurer l'aléatoire et l'incertitude liés aux protocoles d'observation des données réelles et de « modéliser » ainsi le hasard.

C'est au XVII^{e} siècle que, par l'intérêt porté aux jeux de hasard et les problèmes qu'ils génèrent, va se développer le première théorie scientifique et mathématique dite « classique » des probabilités avec Pascal (1623 - 1662), Fermat (1601-1665), J. Bernoulli (1654-1705) et Laplace (1749-1827). Cette première définition « circulaire » qui définit la probabilité d'un événement à partir de l'équiprobabilité des événements élémentaires qui le constituent, donne toutefois pour la première fois le moyen de déterminer numériquement la probabilité d'un événement complexe. Son application reste toutefois limitée aux jeux de hasard, principalement. Plus tardivement, l'approche fréquentielle des probabilités, quant à elle, définit la probabilité d'un événement comme une fréquence théorique, limite (au sens mathématique du terme qui a ses conditions d'existence) de la fréquence relative d'occurrence de l'événement en question, dans une suite de n expériences lorsque n augmente indéfiniment : c'est une approche avant tout empirique (particulièrement fructueuse en physique). Enfin, en 1933, le mathématicien Kolmogorov propose un système probabiliste « décomplété [18]» de trois axiomes, fondé sur la théorie de la mesure. Dans sa généralité, cette approche axiomatique des probabilités englobe les deux premières. Elle définit une probabilité comme une fonction de mesure définie sur un espace abstrait qui représente l'ensemble des événements.

Assortir les conclusions inférentielles, incertaines par nature, d'une certaine mesure du risque d'erreur ou *a contrario* de fiabilité, c'est aussi recourir à la *Théorie des Probabilités.* Toute décision inférentielle relative à une population clairement définie relèvera toujours d'un calcul de probabilité eu égard à la part d'incertitude et d'aléatoire immanente à la démarche inductive en question. Il existe différentes méthodes inférentielles, dont celle bien connue des intervalles de confiance que l'on retrouve dans le traitement des sondages d'opinion. C'est cette méthode qui sera utilisée à titre illustratif dans l'exercice sur la Star Academy qui nous occupe.

Notons que les deux sens du mot « statistique » évoqués plus haut subsistent aujourd'hui. Ils sont cependant liés, par le biais du Calcul des Probabilités, puisque c'est celui-ci qui permet de garantir la crédibilité des mesures statistiques inférentielles obtenues par exemple par des enquêtes, avec les notions de sondages aléatoires et d'intervalles de confiance : ces derniers seront expliqués dans la section suivante.

[18] « Décomplété » au sens du théorème de Gödel.

§2.2 : *D'une Fiction appliquée à la Star Academy :*

Notre sujet de la « Star Academy » permettra d'illustrer par une fiction comment l'Inférence Statistique s'articule avec la Statistique Descriptive et la Théorie des Probabilités, même si le sujet, plus complexe, n'est pas aussi facile à traiter qu'un jeu de Pile ou Face !

En guise d'exercice, et à des fins purement pédagogiques, ciblons une question particulière – et particulièrement simple – qui pourrait se poser au statisticien-probabiliste à propos de la « Star Academy » : *quelle est la proportion (%) de téléspectateurs installés devant le petit écran du samedi soir qui regardaient la « Star Academy » de la dernière saison ?*

Et voici quelle pourrait-être sa réponse…

Il s'agit, avant toute autre chose, de bien définir la « *population* » des téléspectateurs visée car la proportion de téléspectateurs recherchée variera d'une « population » considérée à l'autre. Nous définirons, par exemple, la population ciblée comme étant l'ensemble des téléspectateurs des pays européens ayant une communauté francophone et qui disposent du choix du même ensemble de chaînes de télévision concurrentes (par exemple : TF1, France2, France3, RTL-TVI, Club RTL, TV5 et Canal +).

L'objectif inférentiel est donc de connaître, d'estimer la proportion $\boldsymbol{\pi}$ de la population ainsi définie qui se consacre à la « Star Academy » du samedi soir. Pour savoir « quelque chose » de cette proportion π inconnue de la population, une partie sera prélevée de cette dernière en guise d' « *échantillon* ». Se pose là la question du type d'échantillon à prélever dans le souci d'apporter une information sur la population-mère la plus « valide » possible : il s'agit de collecter l'information avec pertinence. On entend souvent qu'un échantillon doit ainsi être « représentatif » d'une population en ce sens qu'il en constituerait le modèle réduit. La « représentativité » d'un échantillon serait ainsi un argument plaidant en la faveur de sa validité puisque ce « bon » échantillon ressemblerait dès lors autant que faire se peut à la population à étudier, certaines catégories apparaissant en même proportion dans la population et dans l'échantillon. . Y. Tillé (2001) [19] fait remarquer qu'en matière de sondage, « ..cette théorie, couramment véhiculée par les médias et même par certains ouvrages méthodologiques, est erronée : […] pour être valide, un échantillon ne doit pas être représentatif (au sens où nous venons de le définir). Il est, en effet, souvent souhaitable d'effectuer des tirages à probabilités inégales ou de sur-représenter certaines fractions de la population ». De telles techniques d'échantillonnage relèvent du remplacement du tout par la partie aléatoirement choisie de celui-ci : la notion de tirage aléatoire y est centrale. Dans la pratique, rares seront les échantillons purement aléatoires[20], même si la recherche de l'aléatoire dans la constitution de l'échantillon reste au centre des préoccupations pour en assurer la plus grande fiabilité en matière d'inférence.

[19] TILLE Y., *Théorie des Sondages : Echantillonnage et estimation en populations finies*, 2e cycle, Ecole d'Ingénieurs, Dunod, 2001, p. 4-5.

[20] Eu égard à une contrainte de réalité, dans la pratique, on travaille le plus souvent avec des échantillons par quotas, par stratifications, par grappes etc. qui ne sont pas purement aléatoires.

Une fois l'échantillon prélevé, les outils de la *Statistique Descriptive* tenteront d'en obtenir une information résumée la plus exhaustive possible dans le but inférentiel que l'on s'est fixé. La Statistique Descriptive permet certes de gagner en clarté mais engendre forcément une perte d'information qu'un choix adéquat dans la façon de résumer l'échantillon permet d'amenuiser. Pour notre part, la proportion P [21]de l'échantillon, estimateur de la proportion $\boldsymbol{\pi}$ de la population recherchée, résumera de façon exhaustive toute l'information relative à la proportion π de la population, contenue dans l'échantillon. Si sur 500 ménages interrogés, 375 d'entre eux ont déclaré avoir regardé la « Star Academy » le samedi soir, la proportion P de l'échantillon vaudra 75 %. Cette mesure - 75 % de téléspectateurs interrogés déclarent regarder l'émission - ne fait que décrire l'échantillon prélevé ; et absolument rien ne garantit que si le statisticien tirait un nouvel échantillon de 500 nouveaux ménages dans la même population, selon le même mode de prélèvement, on retrouverait la même proportion P égale à 75 %, évidemment !

Pour aller au-delà de la description de notre échantillon et pour savoir s'il y aurait une *probabilité* importante que tout nouvel échantillon de ce type contienne à peu près la même proportion P de téléspectateurs de la « Star Academy », il faut faire une hypothèse d'un modèle probabiliste sur la façon dont les 500 données de notre échantillon ont été générées : dans notre exemple, le statisticien choisirait de faire l'hypothèse stochastique d'une *loi de probabilité* simple (Loi de Bernoulli[22]) où chaque ménage interrogé n'aurait que 2 réponses possibles à donner, génériquement appelées « succès » ou « échec » (« succès » quand le ménage interrogé regarde l'émission et « échec », lorsqu'il ne la regarde pas). Faire l'hypothèse de cette loi de probabilité de Bernoulli réduit effectivement toute l'incertitude de notre problème inférentiel à la seule connaissance du paramètre π recherché, la proportion de la population. Outre cette loi de Bernoulli, la théorie des probabilités offre d'autres lois, comme la loi Normale (ou courbe de Gauss[23], célèbre courbe en cloche) qui permettront, comme dans le cas qui nous préoccupe, d'assortir notre conclusion inférentielle d'un certain niveau de fiabilité.

Parmi les méthodes d'estimation de l'*Inférence Statistique*, il en est une qui permet de construire une fourchette de valeurs dans laquelle on sait qu'il y a une grande probabilité que se trouve la proportion π inconnue de la population: l'estimation par Intervalles de Confiance[24] mesure l'erreur imputable à la démarche inférentielle (erreur d'échantillonnage) qui peut expliquer la distorsion entre la proportion π inconnue de la population et la proportion P observable de l'échantillon.

---

[21] Le lien entre la proportion P de l'échantillon et la proportion π de la population est formellement assurée par le fameuse Loi des Grands Nombres dont il sera question un peu plus loin.

[22] BERNOULLI (Jacques 1er), 1654-1705, mathématicien suisse d'origine anversoise qui compléta le calcul infinitésimal de Leibnitz. Son ouvrage posthume, *Ars Conjectandi*, posa les fondements du calcul des probabilités.

[23] GAUSS (Karl Friedrich), 1777-1855, astronome, physicien et mathématicien allemand, père de la géométrie non-euclidienne.

[24]DESROSIERES A. précise, dans le *Dictionnaire d'Histoire et de Philosophie des Sciences* cité plus haut, que cette méthode d'extrapolation fut introduite par Laplace à la fin du XVIIIe siècle à l'occasion de l'évaluation de la population, dotée d'une « erreur à craindre », ce que les sondeurs modernes appellent l' « intervalle de confiance » ou pour le public, la « fourchette ». Cette méthode fut vivement rejetée dans les années 1820 en raison de la suspicion pour une science de l' « incertain » et de l'approximation. Elle connut une grande notoriété dans les années 1930, quand ces mêmes méthodes d'échantillonnage probabiliste ont été utilisées pour « mesurer l'opinion » ou prévoir les élections (par Georges Gallup aux Etats Unis). Le mot « sondage » en devint synonyme d'enquête d'opinion, alors que son sens premier était celui d'une technique d'échantillonnage de remplacement du tout par la partie aléatoirement choisie de celui-ci.

Plus formellement, cela pourrait s'écrire de la façon suivante :

$$\pi = P \pm \textit{erreur d'échantillonnage}$$

Où cette erreur d'échantillonnage dépendra, entre autres choses, de la taille de l'échantillon sélectionné et du niveau de fiabilité (ou de confiance) que l'on veut pouvoir accorder à cette conclusion inférentielle.

$$\text{Prob}\,(P - \textit{erreur d'échantillonnage} \leq \pi \leq P + \textit{erreur d'échantillonnage}) = 1 - \alpha$$

Cet Intervalle de Confiance permet de « probabiliser » le risque de se tromper ou non : il y a (1-α) % [25]de chance que cette fourchette de valeurs ainsi construite pour π, le contienne effectivement et α % de risque qu'il ne le contienne pas.

Avec notre échantillon de 500 ménages observés, la proportion P observée de l'échantillon égale à 75%, l'hypothèse stochastique d'une loi de Bernoulli et un niveau de fiabilité imposé à 95%, cette fourchette de valeur vaudrait (0.75 – 0.038, 0.75 + 0.038), soit (0.712, 0.788). Il y a donc 95% de chance que la proportion π inconnue de la population se trouve quelque part entre 71% et 79% : il y a par ailleurs 5% de chance que ce même intervalle de valeur construit à partir de notre échantillon ne contienne pas la proportion π inconnue de la population et que cette proportion soit inférieure à 71% ou supérieure à 79%. Et on ne saura jamais, si on a eu la malchance que notre intervalle ne contienne effectivement pas le paramètre π inconnu. Le statisticien-probabiliste fait confiance à ce qui est le plus vraisemblable et ignore les évènements hautement improbables, même s'ils ne sont pas impossibles : c'est exactement là que se situe le risque qu'il prend de se tromper. Cette méthode inférentielle d'estimation par Intervalles de Confiance illustre comment la méthodologie de l'Analyse Statistique appréhende l'incertitude engendrée par la démarche inductive de l'échantillonnage.

Quant au choix intuitif et, certes, guidé par le bon sens, de résumer l'échantillon par la proportion P de l'échantillon en vue d'obtenir une première estimation ponctuelle de la proportion π de la population, il est intéressant de souligner que la célèbre Loi des Grands Nombres (Jacques Bernoulli (1713)) lui apporte une caution théorique.

Cette loi (que l'on peut démontrer) est aux confins des deux courants épistémologiques évoqués plus haut – empirico-inductif et hypothético-déductif – puisqu'elle établit une relation entre la notion de fréquence relative et celle de probabilité. Cette Loi des Grands Nombres a fait couler beaucoup d'encre, depuis Pascal et Leibniz jusqu'aux interprétations statistiques de la physique quantique.
Formellement, cette convergence en probabilité de P vers π de la Loi des grands Nombres peut s'écrire de la façon suivante :

$$\mathbf{P} \underset{n \to \infty}{\overset{p}{\longrightarrow}} \boldsymbol{\pi}$$

Pour notre exercice, elle peut s'interpréter comme nous assurant que lorsque n, le nombre des observations de notre échantillon – observations relatives à la présence ou non devant le petit écran à l'heure de la Star Academy et pour la Star Academy – augmente et

[25] Plus exactement, il y a 100x(1- α) % de chance que cette fourchette de valeurs contienne effectivement π et 100x α % de risque qu'il ne le contienne pas.

augmente encore, on se rapproche de la certitude que la proportion P de l'échantillon se rapproche, quant à elle, encore et encore de la proportion π inconnue et recherchée de la population (d'aussi près que l'on veut , d'une quantité infinitésimal qu'on pourrait nommer ε). P « converge » donc, en probabilité, vers π quand n augmente indéfiniment.

H. Atlan[26] en fait un commentaire très intéressant : ce que nous dit cette Loi, c'est que s'il est impossible de prévoir l'occurrence singulière « dans cet aléatoire moderne où ce qui se produit est *sans cause* connue ni même connaissable, […] la seule prédiction possible est une prédiction statistique qui ne concerne donc plus la survenue d'événements individuels mais celle d'un *ensemble* d'événements que nous construisons en adoptant (pour les besoins de la « cause ») un certain point de vue qui les rend indiscernables ». Il est impossible de prédire si chaque ménage en particulier, dans notre *population,* va ou non regarder notre émission puisqu'on ne connaît pas, dans le détail, la totalité des causes multiples qui détermineront *in fine* si oui ou non, tel ou tel ménage regardera la Star Academy . En revanche, si l'on dispose d'un grand nombre de ménages observés de ce point de vue, il sera possible de prédire avec une grande précision le résultat *en moyenne* - c'est à dire π - si toutefois on considère les ménages observés comme identiques ou indiscernables. Si cette hypothèse de base d' « indiscernabilité » se conçoit sans peine « dans des applications de la thermodynamique statistique où l'on envisage des ensembles pratiquement infinis de molécules identiques […], il est remarquable que des résultats similaires soient obtenus sur des comportements humains de masse, tels que fréquentation d'un lieu public, transports en commun, circulation routière, sondages d'opinion, etc. ». Les causes des comportements individuels de chaque ménage qui le mèneront à regarder ou non l'émission sont multiples et inconnues de l'observateur-statisticien et pourtant le taux d'audience de l'ensemble des téléspectateurs de la Star Academy est « déterminé en ce qu'il obéit à la loi statistique qui le décrit et qui permet de le prévoir ».

§3 Conclusions :

Outre le problème de l'estimation de l'audimat évoqué fictivement pour les besoins de notre exercice, bon nombre d'études pourraient être menées où l'outil statistique et probabiliste serait convoqué . Décrire, sous toutes ses coutures, la structure d'âge du public de la Star Academy ; donner une « photographie de son paysage » pour y faire apparaître des « groupes » de téléspectateurs auxquels seraient associés des « comportements » particuliers ; s'interroger sur le rôle de l'incertitude quant à l'issue de l' « épreuve » comme facteur de fidélisation du public ; prédire les ventes de CD, de magazines et de vêtements en lien avec l'événement télévisuel ; évaluer la probabilité d'être la gagnant ; modéliser le taux d'audience, questionner la causalité du modèle etc.

Suspicion vis-à-vis de la Statistique oblige, on a coutume d'entendre : « Il y a les menteurs, les gros menteurs et puis les statisticiens… ». Evoquons, très succinctement ici, quelques « dérives » de la Statistique qui pourraient cautionner - ou sembler cautionner - cet adage populaire qui lui fait si mauvaise presse !

A ce stade, il n'est pas (trop) difficile de comprendre que cela tient sans doute d'abord du fait que l'on constate que le statisticien se trompe parfois (souvent ?) dans ses prédictions ou estimations. Et c'est sans doute la plus mauvaise critique qu'on puisse lui faire. En effet,

[26] ATLAN H., op.cit, p. 351.

L'Analyse Statistique ne prétend pas au risque zéro ; au contraire, elle « probabilise », mesure même son risque d'erreur, ce qui ne l'empêche évidemment pas de se tromper ! Les méthodes inférentielles de l' Analyse Statistique, à l'instar de celle des Intervalles de Confiance illustrée plus haut, sont là pour donner des *balises* à l'interprétation ; force est de constater que ces balises ne sont que trop souvent galvaudées par des utilisateurs peu scrupuleux et/ou ignorants de cet outil.

Mais il y a aussi, pour tenter de comprendre cette suspicion, les incontournables arbitraires liés à la démarche statistique ; l'arbitraire inhérent au choix du niveau de la fiabilité exigé, celui de la façon d'échantillonner et de résumer un échantillon par une méthode plutôt qu'une autre de la Statistique Descriptive, celui de la méthode inférentielle utilisée, de la « modélisation[27] théorique » choisie et de la construction d'une variable lorsque ce que l'on voudrait mesurer n'est pas disponible, etc. Dans le champs des sciences humaines, c'est un des problèmes majeurs lié à la Statistique ; la difficulté de quantifier ce que l'on veut analyser, soit parce que les données disponibles ne mesurent que très imparfaitement ce que l'on voudrait mesurer, soit parce que ce qui nous préoccupe n'est pas quantifiable, ni même observable[28]. En guise d'illustration de ce côté arbitraire, imaginons que le manager de la Star Academy réclame 80% d'audience pour prendre la décision de relancer une nouvelle saison avec la Star Academy. Strictement parlant, l'intervalle de confiance obtenu à 95% de fiabilité – à savoir (0.712, 0.788) -, l'inciterait à ne pas relancer une telle expérience puisque les 80% exigés ne s'y trouvent pas. Et pourtant, si le statisticien augmentait le taux de confiance à 99% au lieu de 95%, *ceteris paribus*, l'intervalle de confiance serait forcément plus large : (0,70 , 0,80), englobant cette fois, les 80% exigés ! Cependant, les 80% se trouveraient à l'extrémité de l'intervalle ainsi obtenu : si le manager est « statistiquement » averti, il se méfiera d'un tel résultat obtenu aux limites de l'intervalle de confiance remanié.

Autre problème encore ; celui de la qualité d'une théorie en statistique. En effet, la qualité d'une prédiction statistique n'est (même) pas un test suffisant pour juger de la qualité d'une théorie, comme le rappelle Robert Franck[29] : « Si une théorie est finalement rejetée, où se trouve le coupable ? Cette théorie a peut-être été mal spécifiée, les concepts peuvent avoir été mal définis, les indicateurs peuvent être inadéquats, le modèle statistique est peut-être mal choisi, les données sont peut-être déficientes … Il s'avère souvent difficile de localiser le(s) problème(s). En général [le statisticien] ne s'embarrassera pas de la recherche du coupable, et attribuera la faute aux « erreurs de mesures » ! ». C'est la raison pour laquelle l'analyse et la modélisation des erreurs de mesure restera une des préoccupations majeures du statisticien.

L'autre dérive classique souvent dénoncée est celle de l'interprétation causale hâtive dès lors qu'un modèle statistique établit une « relation de dépendance fonctionnelle » entre les grandeurs de deux variables, y = f(x). En guise d'illustration, voici la célèbre fable de l'épidémie de choléra qui aurait sévi sur la Russie des Tsars. Afin d'endiguer l'épidémie, le tsar décida d'envoyer dans toute la Russie infectée un nombre imposant de médecins, et plus particulièrement, dans les villes les plus affectées par la maladie. Un Comité de Santé se

---

[27] ANSPERGER CH. *Formalisation et pertinence dans les sciences sociales,* Chaire Hoover, Réunion de travail aux Facultés Universitaires Saint-Louis, 20 février 2001, p.3. « Construire des modèles et en exploiter les potentialités déductives plus ou moins complexes est l'essence même du travail de modélisation , que ce soit dans la dimension « gratuite » de la compréhension/explication [*comme le fait la Statistique Descriptive*] ou dans la dimension « instrumentale » de la prédiction ».

[28] Dans ce dernier cas cependant, les variables dites *latentes* tentent d'y apporter une solution.

[29] FRANCK R. (1994), *Faut-il trouver aux causes une raison ? L'explication causale dans les sciences humaines*. Publication de l'Institut Interdisciplinaire d'Etudes Epistémologiques, Lyon, p. 38.

penche sur le problème et choisit un certain nombre de villes russes malades. Le Comité y dénombre les médecins et les morts du choléra. Les résultats de l'analyse statistique montre que le nombre de médecins (x) « explique » parfaitement le nombre de morts (y) ; plus précisément, selon une relation fonctionnelle linéaire entre x et y, plus il y aurait de médecins dans une des villes sélectionnées, plus il y aurait de morts du choléra ! Le Tsar décide alors de tuer les médecins…

Cette fable qui montre par l'absurde, les conclusions abusives auxquelles peuvent mener une explication causale sans fondement ne doit cependant pas servir à récuser toute explication causale, quelle qu'elle soit. Comme le souligne R. Franck[30], « l'explication causale ne se heurte pas seulement à des difficultés méthodologiques et épistémologiques ; elle soulève aussi des réticences éthiques, anthropologiques, philosophiques et idéologiques » : on pense souvent qu'appliquer la causalité à l'homme, c'est souscrire au déterminisme, nier la liberté, gommer conscience et raison et ignorer les valeurs de l'esprit. Tout en privilégiant l'étude de l'explication causale, l'auteur dénonce en effet le piège qui consisterait « à s'enfermer dans une alternative qui commanderait de choisir entre un naturalisme qui ne permet pas de rendre raison de la spécificité des phénomènes humains et une quête du sens et des libertés qui, de son côté, refuserait de s'appuyer sur une meilleure connaissance des déterminations où le sens s'inscrit et où nos libertés s'affrontent ». L'explication causale ne se réduit pas à discerner, de x et de y, quelle est la cause, quel est l'effet. L'explication causale demande qu'on explique les lois sur lesquelles le principe d'induction statistique repose ; en effet, le néo-positivisme voulait fonder la science sur le critère de la régularité empirique. En d'autres mots, elle espérait que l'explication scientifique pourrait se suffire de l'exhibition d'un phénomène comme cas particulier d'une loi ; et que la loi elle-même était fondée dès lors que l'expérience attestait que la relation entre variables qui est exprimée dans la loi est régulière, qu'elle se produit toujours ou avec une forte probabilité. L'explication causale ne se limite pas à se demander quelle est la cause du phénomène mais plutôt, pourquoi « y » est régulièrement, fonctionnellement, lié à « x » ; elle consiste à expliquer pourquoi la cause est cause du phénomène !

Plus généralement, à ceux que la formalisation mathématique questionne, on pourrait répondre que toute formalisation, en ce y compris la formalisation mathématique à laquelle recourt la Théorie des Probabilités et l'Analyse Statistique, utilise des concepts, des symboles, etc. qui, par ce biais, tentent d'appréhender « la complexité chaotique du réel »[31] , sans pour autant l'amputer abusivement de ses parties essentielles. La nécessaire « réduction » qu'opère ainsi la démarche statistique ne cherche nullement à nier cette « complexité chaotique du réel » mais plutôt à rendre compte de ses parties essentielles alors qu' « à l'échelle 1 :1, le réel échappe toujours à notre prétention à la connaissance totale[32] ». Sans trop de prétention cependant, ne pourrait-on dire que cette formalisation mathématique *extra-verbale* permet parfois un éclairage, « un type de déduction particulièrement puissant parce qu'il rend parfois possible […] la découverte de conséquences inattendues et/ou inaccessibles par le pur langage verbal[33] ».

Si, dans de nombreux cas, on est tenté de dire, comme Edgard Morrin, qu' il faudrait probablement « computer » un peu moins et « cogiter » un peu plus, dans un grand nombre de

[30] FRANCK R. (1994), op. Cit., *Introduction Générale,* p 1-18.
[31] ANSPERGER CH. op. Cit., p. 3.
[32] ANSPERGER CH. op. Cit., p. 3.
[33] ANSPERGER CH. op. Cit., p. 5.

cas, par contre, ne pourrait-on aussi répondre à l'assertion ci-dessus concernant les statisticiens, qu'il y a, certes, les menteurs, les gros menteurs mais aussi…les « mauvais » statisticiens ?