LA PREUVE PAR LES CHIFFRES (EVIDENCE BASED) : DE QUOI S'AGIT-IL ?

Claudine Schwartz[1]

TITLE

Evidence-based: what is it about?

RÉSUMÉ

La preuve par les chiffres opère dans de nombreux champs disciplinaires. Elle a pour objet la confrontation entre une question de nature théorique et des mesures chiffrées. Nous distinguons trois étapes dans le processus de preuve : une mise en forme de la question, l'usage d'un test statistique et l'argumentation. Nous explicitons le rôle de la rareté dans la procédure de test et aussi ce qui relève d'un consensus social. Nous nous penchons sur la terminologie, notamment celle de « la vraie valeur d'une probabilité (ou d'un paramètre) », et sur la rhétorique employée au 20e siècle qui, en dépit de sa commodité pédagogique, conduisent à une perte de sens de la notion de preuve par les chiffres. D'où une proposition de changement de rhétorique, pour l'adapter à notre époque où la notion de modèle est explicitement présente dans la plupart des démarches scientifiques.

Mots-clés : *preuve statistique, test d'hypothèse, vraie valeur, risque, modèle.*

ABSTRACT

Evidence-based medicine, evidence-based policy, etc.: the statistical proof operates in numerous disciplinary fields. It has for object the confrontation between a theoretical question and quantitative measures. We distinguish three stages in the process of a statistical proof: a shaping of the question, the implementation of a statistical test and the argumentation. We clarify the role of unfrequent events in the procedure of statistical tests and the part which comes from a social consensus. We bend over the terminology, in particular that of "the real value of a probability (or of a parameter)", and over the rhetoric used in the 20th century, which despite its pedagogical convenience, induce a great amount of loss of sense. We suggest a change of rhetoric, to make it cope with the notion of model which actually drives most scientific approaches.

Keywords: *evidence-based, statistical test, true value, risk, model.*

1 Introduction

La notion de preuve par les chiffres, ou *preuve statistique*, a été élaborée entre le 18e et le 20e siècle et mise en forme dans la première moitié du 20e siècle. C'était une époque où les mesures chiffrées constituaient une denrée rare et réservée aux spécialistes. Aujourd'hui, le mouvement « *Open data* » (données ouvertes) conduit à rendre accessible à un large public des fichiers autrefois réservés aux administrations ou aux institutions de recherche. En parallèle, la possibilité de stocker, transmettre, dupliquer et analyser des données, en très grand nombre (les « big data ») transforme profondément le paysage où s'exerce la pensée fondée sur les chiffres. Faire de la statistique au 20e siècle, c'était, entre autre, mener des enquêtes en utilisant des données recueillies selon des plans expérimentaux optimaux ou des

[1] Professeure des Universités, Statisticienne, Paris, claudineschwartz@wanadoo.fr

modes d'échantillonnage réfléchis. A cette pratique s'ajoute maintenant la demande sociétale d'extraire de l'information à partir de fichiers géants et non structurés, obtenus par accumulations de « traces » innombrables et diverses. C'est un peu comme devoir déchiffrer dans des lieux de grand passage un message inconnu, dont on n'est d'ailleurs pas sûr qu'il existe.

La preuve statistique opère dans de nombreux champs disciplinaires. En médecine (*evidence-based medecine*), elle a au siècle dernier contraint les experts (les *mandarins*) à partager le pouvoir d'être « sachant » avec de simples séries de chiffres. L'évaluation des politiques publiques par les chiffres (*evidence based policy*) se développe inexorablement dans le cadre de la démocratie participative et du « new public management ». Des ouvrages majeurs d'Alain Desrosières, tels « La politique des grands nombres » (1993), « Pour une sociologie historique de la quantification » (2008a) ou « Gouverner par les nombres » (2008b), témoignent de la force du sujet et en tracent un cheminement historique. Abhijit Benerjee et Esther Duflo (2011) ont fait de la preuve statistique un outil majeur pour repenser la pauvreté, en confrontant les pratiques sur le terrain avec des schémas théoriques. Dans leur travail, la preuve trouve sa place non comme machine à produire de la « vérité », mais en tant que procédure intégrant en un mouvement continu l'argumentation logique et la prise en compte de données numériques.

Les chiffres ne parlent pas d'eux-mêmes, ce qui fait de la preuve par les chiffres, ou preuve statistique, un passage souvent obligé. Celle-ci doit trouver sa place et sa fonction dans les nouvelles vallées créées par les torrents de chiffres qui naissent un peu partout. Pour contribuer à ce que cette place ne soit pas un ensevelissement, une simple disparition, nous proposons de revisiter l'ensemble du processus de preuve, en explicitant la place de la rareté et celle du consensus social dans la construction des tests statistiques. Nous suggérons aussi de nous dégager de la rhétorique qui a été celle du vingtième siècle pour aller vers un formalisme cohérent avec les pratiques scientifiques actuelles.

Schématiquement, la preuve statistique confronte des objets de nature distincte, à savoir une question théorique et des faits expérimentaux en lien avec celle-ci. Elle œuvre dans les chantiers de la connaissance et produit une parcelle de savoir qui vient d'abord s'inscrire en marge d'un savoir existant et stabilisé. Ou alors elle s'insère dans l'évaluation d'une politique, ou dans l'élaboration d'outils de pilotage.

Des questions simples seront le point de départ pour appréhender le déploiement d'une preuve statistique et ce qu'elle produit en fin de parcours.

2 Naissances : autant de garçons que de filles ?

Cette question a été l'objet de centaines de publications et la source de vives controverses depuis le 17^{e} siècle (cf. Brian et Jaisson, 2007). Elle reste d'actualité dans tous les pays, notamment en Chine avec la masculinisation de la population. Historiquement, elle a aussi servi de banc d'essai pour de nouvelles méthodes en Statistique.

Dans la clinique de la ville où vous résidez, il est né cette année 104 garçons et 97 filles alors que, dans une ville proche, il est né 72 garçons et 62 filles. Si on campe sur une approche purement factuelle, la question de l'égale probabilité d'observer une naissance de fille ou de garçon n'a pas de sens, sauf à admettre des réponses pauvres telles « ça dépend », ou « à peu près ».

Pour s'inviter dans les débats sociaux et scientifiques, la question a dû être conceptualisée. Passons les étapes pour arriver à une formulation utilisant les concepts modernes : « La probabilité p d'observer une naissance de garçon est-elle 1/2 ? »[2]. Ou encore : « Y a-t-il équiprobabilité à la naissance des filles et des garçons ? ». La probabilité est une « chance théorique », nous sommes dans le champ de l'abstraction. Limitons la question dans le temps et l'espace et considérons les 270 844 naissances observées à Paris entre 1901 et 1905 (cf. Hawlbachs et Sauvy, 1936), dont on ne retient que le sexe. Tout est en place pour l'administration d'une preuve par les chiffres, dans ce cas très simple où l'ensemble des données à recueillir ne nécessite pas de définir la mesure à faire et la construction d'un plan d'expérience ou d'un plan d'échantillonnage (stratification, cohorte, échantillon randomisé, etc.).

La phase suivante est la mise en œuvre d'un test statistique. Pour cela, on associe à la question une hypothèse, qui sera ici : les sexes à la naissance sont équiprobables. Nous reviendrons ultérieurement sur ce choix précis d'hypothèse. Un test est une procédure codifiée de confrontation entre deux objets de nature bien distincte, une hypothèse théorique et des données numériques en lien avec cette hypothèse. Le résultat de cette confrontation sera de dire que l'hypothèse théorique est compatible ou non avec les données. On spécifie complètement la procédure avant de connaître les données, ce qui permet, quand on en dispose, de répondre automatiquement par oui (c'est compatible) ou non (ce n'est pas compatible).

Une procédure spontanée de confrontation serait ici de conclure positivement à la compatibilité si la série de données conduit à un taux de masculinité (proportion de garçons) *voisin* de 0,5, en spécifiant ce qu'on entend par *voisin*. On pourrait par exemple choisir que c'est le cas pour un taux de masculinité situé entre 0,49 et 0,51, c'est-à-dire qui s'écarte de la valeur théorique 1/2 de moins de 1% (mais on pourrait aussi choisir 5% ou 0,1%, un tel choix est arbitraire). Avec cette procédure spontanée, on dirait que les données de la France entre 1900 et 1905 sont compatibles avec l'équiprobabilité. En effet, le taux de masculinité calculé sur les $N = 270\,844$ naissances est $f = 0{,}508$ (soit une proportion de naissances de garçons égale à 50,8%). Cette procédure est fondée sur l'intuition juste que, sous l'hypothèse d'équiprobabilité, la fréquence doit être proche de 1/2, le nombre 1/2 étant l'invariant théorique autour duquel fluctuent les fréquences observées. Elle n'est néanmoins pas reconnue car elle ne tient aucun compte de l'état actuel de la science, et, plus précisément, des théorèmes démontrant que sous l'hypothèse d'équiprobabilité, plus on a de données, *plus* le taux de masculinité est proche de 1/2 : un écart de 1% sur 100 données ou 10 000 données, c'est très différent. Ces théorèmes permettent fort heureusement de quantifier ce *plus*. Ainsi, une fréquence de 0,508 calculée sur 1 000 données ou sur 270 844 données ne s'interprète pas du tout de la même manière. Une donnée brute, même aussi simple qu'un taux de masculinité, ne parle pas d'elle-même.

La notion de preuve *par les chiffres* repose sur un consensus fort : les données expérimentales y ont le dessus sur l'hypothèse théorique. Autrement dit, si les chiffres peuvent contredire l'hypothèse théorique, l'inverse est exclu.

Parler de contradiction entre une hypothèse théorique et des données (*data*) est simple à concevoir dans le cadre abstrait de la logique mathématique. Ainsi, si l'hypothèse est que tous

[2] Cette probabilité n'est pas celle de concevoir un garçon : parmi les avortements, spontanés ou non, les proportions de filles et de garçons ne sont pas égales. La probabilité de concevoir un garçon est un paramètre biologique, que nous n'aborderons pas ici.

les moutons sont blancs, un unique mouton noir est à lui seul une contradiction. Mais comment un taux de masculinité observé peut-il contredire une hypothèse d'équiprobabilité ? En effet, dans le cas d'un automate fabriquant une liste de symboles F ou G avec équiprobabilité et indépendance des choix, la fréquence 0 (ou 1a fréquence 1) du symbole G est possible. Possible, oui, mais rare et évidemment d'autant plus rare que la liste est longue ! La notion de rareté va permettre d'élargir la notion intuitive de contradiction. Nous emploierons le terme d'incompatibilité et non plus de contradiction pour éviter toute confusion. Parler de contradiction est d'ailleurs une commodité de langage trompeuse, en ce qu'elle engage implicitement la pensée dans un cadre booléen : la contradiction est ou n'est pas, c'est oui ou c'est non. L'état de *rareté* se situe par contre dans un continuum et il faut quantifier le « combien rare ».

Le principe d'un test statistique est le suivant : si les résultats présentent un certain *caractère, rare sous l'hypothèse testée,* défini avant de les connaître, alors ils sont déclarés incompatibles avec l'hypothèse. Si les résultats ne présentent pas ce caractère, alors ils sont déclarés compatibles avec cette hypothèse. Ou encore, l'étude de la compatibilité entre une hypothèse théorique et des données repose sur la construction mathématique, *a priori*, d'une *région de rareté* (sous l'hypothèse à tester), région composée de tous les résultats (les données) qui ont ce caractère. Ce dernier est défini mathématiquement pour optimiser certaines propriétés du test, lui conférant ainsi des qualités spécifiques. Par abus de langage, nous qualifierons de rares des résultats (ou les valeurs correspondantes d'un ou plusieurs indicateurs[3] construits à partir d'eux) qui sont dans la *région de rareté*. Nous dirons que les résultats sont communs sinon.

Dans le cas des naissances à Paris entre 1901 et 1905, les résultats sont les éléments d'une liste de lettres F ou G qui codent le sexe des naissances, le seul indicateur associé au test étant ici le taux de masculinité (proportion de G dans la liste). On sépare les valeurs possibles de ce taux en deux lots : celles qui sont *rares* en cas d'équiprobabilité et les autres valeurs, dites *communes*[4]. La notion de rareté est quantifiée par un nombre appelé risque α (en pratique plus petit que 1/10) : son choix est arbitraire même si dans certaines communautés, un consensus s'est établi sur l'usage d'une valeur particulière (ainsi 0,05 en médecine). Ce risque est la probabilité, pour une liste de symboles F et G fabriqués par un automate à partir de l'équiprobabilité, de produire un taux *rare* de symbole G. Appliquer un test statistique à des données fournies par un automate consiste à étalonner ce test dans des conditions « idéales », α étant alors le risque de rejeter à tort l'hypothèse testée.

Choisissons $\alpha = 0{,}001$. Le résultat du test est négatif, il n'y a pas compatibilité entre l'équiprobabilité et les données du début du 20^{e} siècle, l'hypothèse d'équiprobabilité est rejetée. L'écart entre la fréquence 0,508 observée est dit significatif au risque 0,1%. Cet écart, égal à 0,008, peut sembler faible, mais il est calculé sur un grand nombre de données.

Le choix de α quantifie la rareté de l'ensemble qui conduit au rejet de l'hypothèse testée lorsque les données sont produites par un automate. Prendre une valeur faible (0,00001 par exemple) conduirait en pratique à considérer facilement que l'hypothèse est compatible avec

[3] Ces indicateurs composent ce qu'on appelle la statistique du test

[4] On notera que, dans le vocabulaire utilisé, l'adjectif *commun* appliqué à un taux de masculinité est hérité de sa seule appartenance à un ensemble dit *commun* et ne signifie pas pour autant que cette valeur particulière du taux est fréquente. En fait, en cas de données en grand nombre, toutes les valeurs possibles du taux de masculinité, prises individuellement, sont de faible probabilité. Celles qui sont déclarées *communes* sont celles qui se situent dans la plage « attendue » (commune) des valeurs sous l'hypothèse testée.

les données mais la validation ainsi produite de l'hypothèse testée serait faible. Par contre, si la confrontation est négative, une faible valeur de α augmente la puissance du résultat obtenu. Comme on ne sait à l'avance quel sera le résultat du test et qu'il convient de choisir α avant de le faire, il y a un compromis à trouver[5]. La valeur 0,05 est standard depuis l'origine de ce type de test (R. Fischer, 1925), au point que c'est une valeur par défaut qu'on omet souvent de mentionner. Une des raisons de ce choix pourrait être qu'il conduit à des calculs simples, puisque l'ensemble des valeurs communes du taux de masculinité sous l'hypothèse d'équiprobabilité est l'intervalle $[1/2 - 1/\sqrt{N},\ 1/2 + 1/\sqrt{N}]$. Les autres valeurs classiquement utilisées pour α sont 0,01 ou 0,001. Des valeurs plus faibles ne sont envisageables que si on a beaucoup de données, mais se posent alors éventuellement d'autres problèmes de précision du modèle testé.

Après le temps de la conceptualisation et de la formalisation d'hypothèses, et celui du recueil des données, après le temps de la mise en œuvre d'un test, la dernière phase du déroulement d'une preuve statistique sera le temps de la pensée argumentative, de l'inscription du résultat du test dans le cadre théorique ou sociétal où est née la question. Nous l'abordons maintenant.

Le résultat du test sur les naissances constitue un *élément de preuve* (*piece of evidence* en anglais) de la non-équiprobabilité des naissances à Paris au début du 20^{e} siècle. Il est conséquent, vu le nombre des données et la valeur du risque α, ici égal à 0,001. Comme de très nombreuses autres séries de données recueillies sur toute la France et à d'autres époques ainsi que dans toute l'Europe ont toujours conduit à la même conclusion, avec un taux de masculinité toujours supérieur à 0,5, les *éléments de preuve* sont venus se renforcer les uns les autres. Prenant acte de tous, on considère aujourd'hui qu'il n'y a pas équiprobabilité des sexes à la naissance et qu'en moyenne, il naît plus de garçons. On n'en réfère plus à des données, on ne parle plus de risque α, la connaissance est solidifiée (ce qui ne signifie pas « incassable »). La rareté qui a servi tout au long du processus de preuve a disparu.

Si la probabilité d'observer une naissance de garçon n'est pas 1/2, peut-on lui attribuer une valeur théorique autour de laquelle fluctuent les fréquences observées en des lieux et des temps différents ? Ou alors, d'une décennie à l'autre, d'un pays à l'autre, le taux de masculinité change si radicalement que ce modèle probabiliste simple ne serait pas pertinent ? La question est ancienne. En 1711, John Arbuthnot, mathématicien et médecin écossais a « montré » que la divine providence avait maintenu un taux de masculinité constant à Londres entre 1629 et 1710. Trente ans plus tard, le pasteur et démographe prussien Peter Süssmilch, dans son ouvrage *L'Ordre Divin* (*Die göttliche Ordnung*) paru en 1741, arrive à la même conclusion, et de plus il chiffre la constante : il naît 105 garçons pour 100 filles, soit un taux de masculinité à la naissance égal à 105/205. En 1814, le marquis Pierre-Simon de Laplace publie un jeu de données relatives aux naissances des années 1800, 1801 et 1802 dans

[5] Le choix d'une valeur de α introduit un couperet artificiel, notamment si les données passent juste la barre de ce seuil, ou sont justes en deçà. Les ordinateurs permettent à présent le calcul d'une probabilité, appelée *p*-valeur, à laquelle l'utilisateur peut comparer le risque α qu'il avait choisi : si α est supérieur (resp. inférieur) à cette *p*-valeur, les données sont non compatibles (resp. compatibles) avec l'hypothèse testée. Cette *p*-valeur, calculée quand les données sont connues, est délicate à interpréter. Aussi, en pratique, on réintroduit plusieurs seuils en décorant le résultat du test par une étoile si la *p*-valeur associée est entre 0,05 et 0,01, deux étoiles si elle est entre 0,01 et 0,001 et trois étoiles si elle est inférieure à 0,001. Ces étoiles permettent de visualiser la force de la significativité d'un écart observé (plus il a d'étoiles, plus il est significatif). Ce n'est pas là une solution qui règle l'arbitraire du choix de α, mais quand on a de l'expérience dans le traitement des données, la *p*-valeur ou les étoiles sont des outils utiles.

certaines communes de France et trouve un taux de masculinité égal à 0,5116 (110 312 garçons et 105 287 filles). La valeur théorique de référence actuelle est $p = 105/205 \sim 0{,}5122$ (on dit souvent qu'il naît 105 garçons pour 100 filles). Vérifions sa compatibilité avec des données actuelles. En France métropolitaine, le taux de masculinité, calculé sur les $N = 7\,810\,779$ naissances entre 2001 et 2010[6] est $f = 0{,}5117$ (soit une proportion de naissances de garçons égale à 51,17%). La procédure de confrontation entre cette valeur et les données conduit à un résultat positif : il y a compatibilité entre la valeur $p = 105/205$ et les données du début du 21e siècle. L'écart entre la fréquence observée et la probabilité de référence n'est pas *significatif,* il est *commun*.

Chaque pays définit à partir de ses propres données une valeur de référence (ou valeur théorique) pour la probabilité d'observer une naissance d'un garçon, ce qui permet ensuite de tester la significativité de dérives éventuelles et d'en chercher des explications. La probabilité de référence en Asie vers les années 1981 était environ 0,517 (107 garçons pour 100 filles) (cf. Attané, 2008). Le taux de masculinité de la population de la Chine, observé en 2005, vaut 0,546 (120,5 garçons pour 100 filles) et est significativement différent de 0,517, avec $\alpha = 0{,}001$, ce qui justifie d'en rechercher des causes. L'avortement sélectif, l'infanticide des filles à la naissance (la naissance de celle-ci étant alors non déclarée) sont au rang des explications avancées et confirmées par d'autres données. Selon les Nations Unies, on serait redescendu à 108 garçons pour 100 filles en 2010, mais avec une grande variabilité suivant les régions.

3 Télépathie

La télépathie traite de l'échange d'information entre deux personnes en dehors de toute transmission sensorielle. De nombreuses expériences ont été pratiquées sur ce sujet, proches de celle que nous imaginons ci-dessous. Dans cette expérience fictive, Paul, dont on veut tester les dons de télépathie, est dans une sorte de bunker, sans moyen aucun de communication avec l'extérieur. On fait défiler devant Jeanne des chiffres 0 ou 1, les choix étant équiprobables et indépendants. Jeanne doit les regarder, y penser un certain temps et Paul doit alors noter le chiffre auquel il croit que Jeanne pense. La réponse est codée S comme succès si le chiffre noté par Paul est celui qui a défilé sur l'écran devant Jeanne, et E comme échec sinon. On fait 100 fois cette expérience (en plusieurs jours pour éviter la fatigue) et on calcule la fréquence f de S dans la liste des 100 symboles S ou E obtenus. Que faire de cette fréquence ? A partir de quand estimera-t-on qu'il y a télépathie ?

Si Paul a vu juste dans 53% des cas, va-t-on déclarer qu'il est télépathe et que la télépathie est prouvée ? Comparer le taux de réussite à 0 n'aurait pas de sens car se tromper systématiquement serait aussi extraordinaire que ne jamais se tromper. On pourrait comparer les résultats de Paul avec ceux d'un automate qui produit une liste de symboles 0 ou 1 suivant une procédure à définir (choix de 1 avec une probabilité p ou alterner régulièrement des 0 et des 1 ou répondre toujours 1 ou toujours 0, etc.) : si la réponse de l'automate est celle qui a été présentée à Jeanne, on code S et sinon E. Le choix de la stratégie d'un automate est complètement arbitraire, mais fort heureusement n'a aucune incidence sur sa probabilité de succès : on démontre que, quelle qu'elle soit, compte tenu du mode de production des chiffres 0 ou 1 (équiprobabilité et indépendance des choix) la probabilité de bonne réponse est 1/2. Il

[6] Données de l'INSEE, statistiques de l'état civil.

suffit donc finalement de confronter directement la fréquence de succès de Paul à la probabilité 1/2 (et non à la probabilité 0 comme certains le pensent de prime abord). On est dans la même situation que précédemment, à savoir regarder la compatibilité d'une fréquence avec une probabilité, ici 1/2.

Deux issues au test sont possibles.

Situation 1 : l'écart entre la fréquence de S observée et 1/2 est significatif au risque α choisi avant l'expérience. A-t-on pour autant « statistiquement démontré » que la télépathie existe ? Evidemment non, la confrontation d'un modèle (ici l'équiprobabilité) à une unique série de données ne peut pas à elle seule « démontrer » l'existence d'un phénomène physique.

Un tel résultat constituerait un élément de preuve, dont la puissance est liée au choix de α et au nombre de données dont on dispose. Pour que la télépathie entre dans le champ des connaissances scientifiques, il faudrait une accumulation importante d'éléments de preuve, certains venant de ce type d'expériences, d'autres liés par exemple à l'analyse d'images cérébrales et d'autres encore à l'élaboration d'un modèle physiologique. On peut surtout penser qu'il faudrait rattacher la notion de télépathie à d'autres concepts et, pourquoi pas, intégrer dans ce sujet les expériences de psychologie cognitive telles que celles de Soon *et al.* (2008) ou de Haynes et Rees (2006) où un ordinateur devine ce qu'un individu va faire avant même qu'il y pense consciemment.

Dans un autre registre, le lien causal entre tabac et certains cancers fait partie du savoir scientifique aujourd'hui établi (solidifié). Cela s'est fait par accumulation de dizaines d'éléments de preuves : des preuves par des mesures sur des humains ou sur des animaux, d'autres étant déduites de modèles ou de théories déjà acceptées.

Situation 2 : Paul ne se distingue pas, quant à ses succès de divination, de ceux qui sont produits en répondant n'importe comment (au hasard ou selon une règle précise). Au niveau de l'expérience faite, l'éventuel don de télépathie de Paul n'est pas perceptible.

La preuve directe de non existence d'un phénomène est beaucoup plus délicate que celle de son existence. Il ne s'agit pas là d'un problème propre à la statistique. Paul ne serait sans doute pas ébranlé par cette expérience, et pour de bonnes raisons. Il pourrait argumenter que ce cadre rigoureux le perturbe. Il pourrait prendre acte que cette expérience indique que ses dons ne s'exercent pas sur des suites de 0 ou de 1 choisies au hasard, mais qu'elle ne dit rien de ses dons de télépathie pour des pensées qui ont du sens.

Et il pourrait continuer à faire des expériences, de plus en plus sophistiquées ; un tel processus est sans fin.

Par contre, les sceptiques vis-à-vis de la télépathie (cf. en référence l'adresse du site « Les Sceptiques du Québec »), eux, seraient confortés dans leur idée *a priori* et peu motivés pour mettre en œuvre de nouvelles expériences.

La télépathie est un sujet ésotérique, mais néanmoins représentatif de bien d'autres : l'hypothèse testée n'est pas reliée à un corpus théorique, elle est en dehors des connaissances actuelles. C'est une question parmi des milliers d'autres, venant de tous les domaines, qui donnent lieu à des éléments de preuves dont la plupart resteront à jamais flottants, en attente, tandis que quelques-uns seront rattachés à un corpus de connaissances solidifiées.

Un exemple célèbre est celui de l'effet de l'absorption de fortes doses de vitamine C pour prévenir ou guérir le rhume. On trouvera une bibliographie bien documentée des travaux entrepris entre 1970 et 2011 concernant ce sujet sur le site de Hemilä (2011). Il s'agit *a priori*

de quelque chose de beaucoup plus simple que de montrer l'existence d'un phénomène qu'on appellerait la télépathie, mais voyons de plus près. En 1970, Linus Pauling, prix Nobel de chimie en 1954 et prix Nobel de la paix en 1962, publie un best-seller « Vitamin C and the Common Cold » dans lequel il fait part de sa conviction de l'effet bénéfique de la vitamine C et des « preuves » qui la sous-tendent. Compte tenu de la notoriété de l'auteur, la vente de vitamine C a connu un pic spectaculaire aux USA, la *Food and Drug Administration* s'en est émue et de nombreux essais thérapeutiques ont été entrepris. Aujourd'hui, on parle souvent, en feuilletant Internet sur ce sujet, de *croyance populaire*, terme approprié puisqu'il ne dit rien de preuves éventuelles. En 2009 est parue une analyse conjointe d'une trentaine d'essais thérapeutiques, enregistrés par la fondation Cochrane spécialisée dans l'analyse des essais sur un même sujet (Hemilä *et al.*, 2009). Cette étude conclut que les données sur la prise de fortes doses de vitamine C n'ont pas fait apparaître d'efficacité pour la prévention ou la guérison, et celle-ci n'est donc pas « rationnellement » justifiée pour toute une population, bien que des éléments de preuve aillent en sens contraire dans le cas d'individus soumis à un entraînement sportif intense ou à un environnent très froid[7]. Il semblerait que cette méta-analyse de 2009 ait clos le débat… au moins provisoirement.

4 Naissances légitimes et illégitimes en France en 1824-1825

L'étude des variations du taux de masculinité suivant les régions du globe, les époques, la différence d'âge des parents, leur origine socioculturelle et toute sorte d'autres critères ont été publiées dans des revues savantes. Nous avons choisi ici l'étude de la légitimité ou de l'illégitimité des naissances pour illustrer certains aspects de la notion de preuve statistique.

En 1840, Jules Gavarret, médecin français, pose quelques-unes des premières pierres de l'épidémiologie. Parmi les nombreux exemples qu'il traite, on trouve des données issues de relevés publiés par le ministre de l'intérieur, qui ont permis à l'auteur de dresser le tableau 1 ci-dessous et de calculer les taux de masculinité f et f' pour les naissances légitimes et pour les naissances illégitimes en France pendant les années 1824 et 1825.

TABLEAU 1 – *Naissances en 1824-1825 selon le sexe et la légitimité*

	Naissances légitimes	Naissances illégitimes
Garçons	939 641	71 661
Filles	877 931	68 905
Total	1 817 572	140 566
Taux de masculinité	f = 0,5170	f' = 0,5098

Construire le tableau ci-dessus, c'est-à-dire croiser le sexe à la naissance et la légitimité de celle-ci, n'est pas neutre et témoigne des préoccupations du 19e siècle. Les facteurs ne se croisent pas spontanément, tout tableau ainsi construit porte l'idée *a priori* qu'il pourrait y avoir un lien et que celui-ci fait sens.

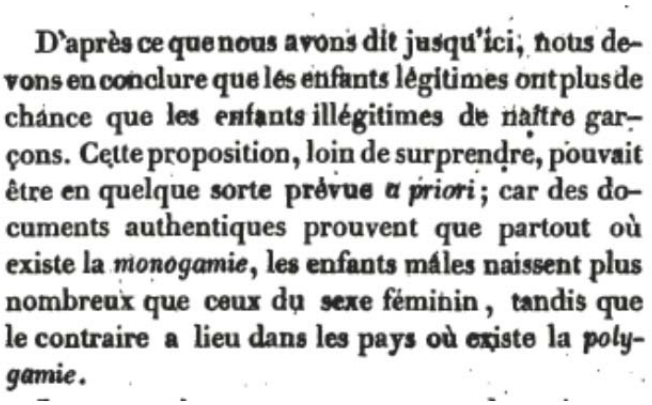
La question théorique à laquelle J. Gavarret confronte les deux taux de masculinité observés pourrait s'énoncer : *les probabilités de naître garçon sont-elles les mêmes pour les naissances légitimes ou illégitimes ?*

[7] On notera que Linus Pauling e été un des premiers à ne pas se satisfaire d'une seule étude et qu'il a publié dans les *Proceedings of National Academy of* Sciences un des premiers articles de ce qui deviendra la meta-analyse (Pauling, 1971).

Il ne s'agit pas ici de démontrer à l'aide de mesures expérimentales que deux nombres (la probabilité de naître garçon et 1/2) sont strictement égaux. La preuve d'une telle égalité, concernant deux objets abstraits (des nombres) ne pourrait d'ailleurs être envisagée que dans le champ de la théorie. Le sens de cette formulation est en fait : *existe-il une valeur de la probabilité de naître garçon compatible avec les deux séries de données ?* La probabilité p de naître garçon est le paramètre d'un modèle qui rend compte de l'aléa du sexe d'un enfant à venir, et plusieurs modèles, pour des valeurs voisines de p, sont compatibles avec un même ensemble de données, quel que soit sa taille.

La procédure utilisée pour confronter les données du tableau ci-dessus à l'hypothèse d'égales probabilités de naître garçon suivant la légitimité ou non de la naissance est là aussi un test statistique. Le résultat de ce test dit que les données ne relèvent pas d'un même modèle. L'écart entre les fréquences de garçons (taux de masculinité à la naissance) est de 0,007 (soit 0,7%). Bien que faible, cette différence, recueillie sur des données nombreuses, s'avère significative au risque $\alpha = 0{,}001$. On dit aussi qu'il y a un lien statistique significatif au seuil $\alpha = 0{,}001$ entre les deux caractères étudiés, légitimité et sexe.

La suite du déroulement de la preuve par les chiffres est la phase d'interprétation du lien statistique. Jules Gavarret disposait d'une procédure de test voisine de celle qu'on appliquerait aujourd'hui. Son interprétation est donnée dans la figure 1.



D'après ce que nous avons dit jusqu'ici, nous devons en conclure que les enfants légitimes ont plus de chance que les enfants illégitimes de naître garçons. Cette proposition, loin de surprendre, pouvait être en quelque sorte prévue *a priori* ; car des documents authentiques prouvent que partout où existe la *monogamie*, les enfants mâles naissent plus nombreux que ceux du sexe féminin, tandis que le contraire a lieu dans les pays où existe la *polygamie*.

FIGURE 1 – *Interprétation de Jules Gavarret (en 1840)*

Les controverses, comme on peut l'imaginer, sont allées bon train, les *documents authentiques prouvant* le pouvoir explicatif de la polygamie n'ayant pas été jugés convaincants par tous. Montesquieu, dans *l'Esprit des lois*, publié en 1848, exprime son refus de toute interprétation directement causale entre ce qui relève du registre moral ou religieux (un ordre divin, la Providence) et ce qui relève du biologique (le sexe à la naissance). Selon lui[8], les climats chauds provoqueraient une légère baisse du taux de masculinité. D'où un excès de femmes qui rendrait nécessaire la polygamie afin que chacune d'elle puisse participer à l'effort reproductif de l'espèce humaine. Autrement dit, le taux de masculinité faible serait la cause et non l'effet de la polygamie.

Reprenons les données de Gavarret, à la lumière du 21^{e} siècle. Le résultat de la confrontation ne nous paraît pas prévisible. Comme Montesquieu, nous n'admettons pas que le lien statistique entre légitimité de la conception et taux de masculinité soit un lien causal direct. Ne croyant pas que les lois de la biologie se conforment à la morale d'une époque,

[8] Le titre originel du chapitre qui traite de ce sujet était « Comment la Loi de Polygamie est une affaire de calcul ».

comment alors interpréter ce résultat ? On pourrait évidemment dire que les données sont mal recueillies. Ce ne serait pas simplement être mauvais joueur. En effet, on ne connaît pas le détail du mode de comptage et celui-ci, comme toute mesure, a sa part d'incertitude. Les chiffres sont donnés à l'unité près dans le tableau 1, mais la précision du comptage, inconnue, n'est sûrement pas l'unité.

Si le lien statistique observé ci-dessus avait été isolé, n'avait été observé que sur ces seules données, on l'aurait vite oublié. En effet, si on regarde le taux de masculinité sous tous les angles, en fonction de très nombreux critères, on trouvera que chaque époque offre des résultats « significativement exceptionnels » dont on ne saura jamais s'ils témoignent d'une fluctuation rare ou d'une cause spécifique. Mais ici, plusieurs études de cette époque ont confirmé le lien statistique. On peut regretter l'orientation de ce type d'étude, mais nier le lien statistique n'est pas pour autant une explication admissible.

L'interprétation de Montesquieu concernant l'influence du climat sur le taux de masculinité à la naissance n'a pas été confirmée par des données ultérieures. Une autre explication serait la suivante. Il y aurait eu à l'époque où les données de Gavarret ont été recueillies un taux plus élevé d'enfants mort-nés ou de grossesses non menées à terme dans le cas d'illégitimité (conditions de vie plus difficiles) ; on sait par ailleurs que le taux de masculinité est plus élevé chez les enfants mort-nés ou pour des grossesses non menées à terme (au niveau des derniers mois)[9]. Ainsi, en cas de grossesse illégitime, les possibilités de naissances de garçons disparaîtraient plus que celles de filles. Une autre explication avancée est qu'au moment de la naissance, les garçons auraient été plus facilement reconnus (légitimés) que les filles. On ne peut pas remonter le temps et recueillir de nouvelles données pour mettre ces explications à l'épreuve des faits expérimentaux. Nous les indiquons simplement pour montrer que plusieurs explications sont possibles, qui peuvent d'ailleurs cohabiter.

Ne pas interpréter un lien ou une corrélation statistique en termes de causalité directe est un slogan que les enseignants de Statistique martèlent tout au long de leur carrière. Malgré tout, la conviction très répandue que les chiffres parlent d'eux-mêmes opère en sens contraire, et on est donc en France assez précautionneux à ce sujet. Ainsi, en dehors d'enquêtes sur les discriminations raciales, on estime que croiser des données ethniques avec d'autres données risque de stigmatiser des groupes définis par un critère non explicatif. Aux Etats-Unis, la situation est toute autre. Ainsi, dans le cadre de l'Open Data, le FBI a ouvert il y a peu de temps son fichier sur les homicides aux USA entre 2000 et 2010 (165 068 homicides), que le *Wall Street Journal* a interfacé (cf. adresse du site en référence). Les variables renseignées pour presque chaque homicide, en dehors du lieu et de l'année, sont la race et le sexe de la victime et de l'assassin, le lien éventuel qui les unit, le mode d'action. L'interface du *Wall Street Journal* (voir figure 2) présente alors, en tout premier lieu, pour des valeurs choisies de ces facteurs, les diagrammes circulaires donnant la race de la victime et de l'assassin, ce qui serait inconcevable en France. L'interface du *Wall Street Journal* est un outil permettant une visite semi-guidée des données, qui démocratise la possibilité d'élaborer un questionnement. Cependant, construire avec ce type d'outils une image cohérente et un peu subtile de la criminalité aux USA, impliquant éventuellement d'engager un processus de preuve pour certaines questions, ne va pas de soi. La technique est actuellement en avance sur la réflexion

[9] Les taux de masculinité des enfants mort-nés varient de 56% à 59% dans le monde (cf. Hawlbachs et Sauvy, 1936).

quant à l'usage des données chiffrées, mais la nécessité d'une telle réflexion va se faire impérieuse, personne ne souhaitant le règne d'une réflexion robotisée.

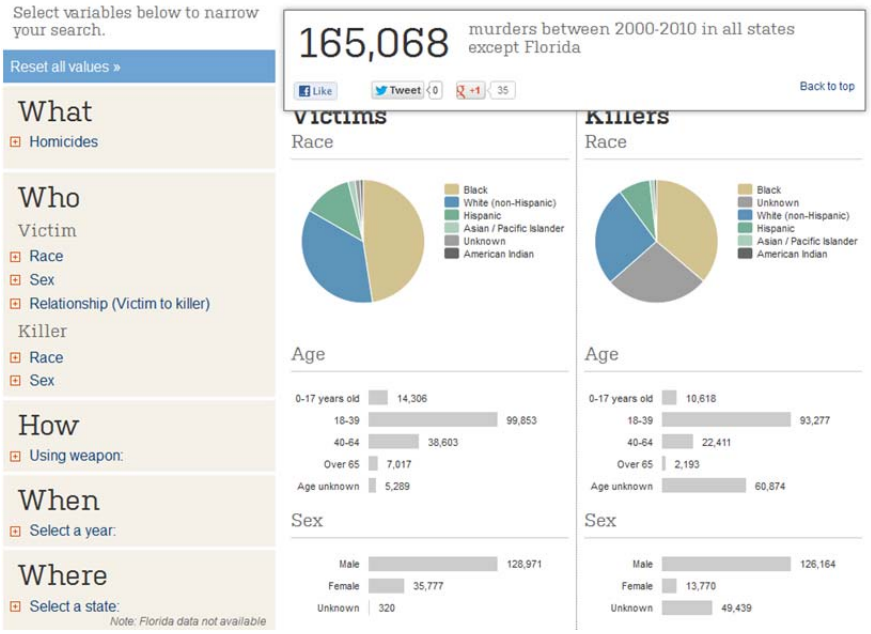


FIGURE 2 – *La page d'accueil de l'interface proposée par le Wall Street journal*

En France, le gouvernement a aussi ouvert ses fichiers et l'observatoire de la délinquance a produit des cartes (voir figure 3) ; l'intérêt se porte sur la localisation des événements et il semblerait qu'on n'ait pas d'autres données librement accessibles (cf. en référence le site Cartocrime).

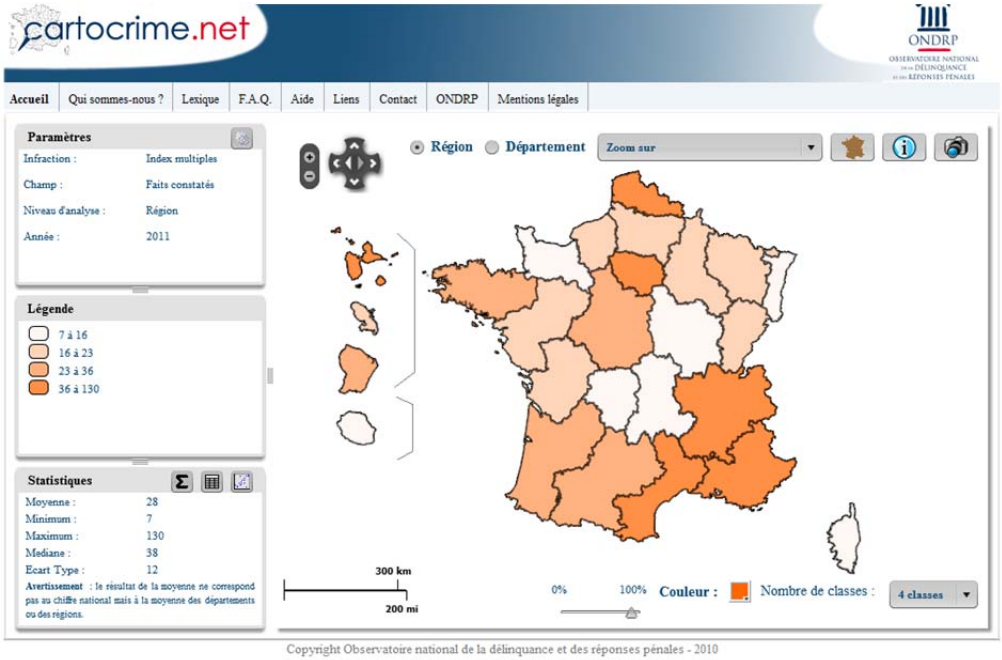


FIGURE 3 – *Carte par départements des lieux des homicides en France en 2011. Les données sont accessibles en tapant sur le logo « tableau » correspondant sur l'écran*

Dans les exemples ci-dessus, on analyse des données classées suivant deux critères, l'étude n'étant pas *a priori* liée à une décision. Des situations équivalentes au plan de la statistique peuvent aussi relever de motivations autres, par exemple lorsque les données sont liées à une action, dans le cadre d'une étude de faisabilité ou d'une évaluation (*evidence-based policy*). Dans leur ouvrage sur la pauvreté, Banerjee et Duflo (2011) en donnent de très nombreux exemples. Ainsi, devant l'échec des campagnes de vaccination des nourrissons dans la région d'Udaipur en Inde, on a envisagé de donner un kilo de lentilles aux parents pour chaque séance de vaccin, plus un cadeau à la fin de la série des vaccins. On peut argumenter bien longtemps sur les nombreux déterminants susceptibles d'intervenir dans la décision des parents[10] et qui auraient partie liée à l'efficacité ou l'inefficacité de cette action. Les auteurs ont choisi d'expérimenter : 30 villages ont été choisis au hasard parmi soixante, et leurs habitants ont reçu les cadeaux. On a comparé la proportion de vaccinations dans l'ensemble des villages choisis avec celle des autres villages. La différence en faveur des villages où a été mise en œuvre la campagne de cadeaux a été statistiquement significative (et même bien au-delà de ce qui était espéré). Il est même apparu que les villages voisins de ceux qui recevaient des cadeaux se vaccinaient plus qu'avant l'expérimentation. La preuve par les chiffres a convaincu les décideurs de mettre en œuvre dans la région cette campagne de cadeaux, même si les taux les plus élevés atteints dans certains villages (38%) étaient loin d'être satisfaisants. Cette action pourra bien sûr être réévaluée dans la durée. La preuve statistique ne visait pas à produire un savoir autre que local dans le temps et l'espace, mais elle a été décisive dans l'étude de faisabilité de l'action envisagée.

5 La rhétorique de la preuve

Les tests statistiques font aujourd'hui l'objet d'une présentation codifiée et d'une rhétorique imposée. Nous allons faire le lien entre les termes employés ci-dessus et ceux des livres d'enseignement, en remontant à quelques sources historiques de la théorie des tests.

Deux types de situations conduisent à l'usage de tests statistiques : la preuve d'une idée (ou d'un concept) par les chiffres d'une part, la validation de modèle d'autre part.

5.1 La preuve d'une idée par les chiffres

Partant d'une question, par exemple, « *Naît-il autant de garçons que de filles ?* », on envisage, pour la résoudre, une famille de modèles (ici des modèles binomiaux) qui permettront de formaliser la question en termes de paramètre de modèle. Ce paramètre est ici la probabilité, notons la p, d'observer une naissance de garçon et la question posée est codée sous la forme d'une hypothèse, dite hypothèse nulle H_0. Nous noterons « $H_0 : p = 1/2$ ». Pour la comparaison des taux de masculinité en cas de naissances légitimes ou illégitimes, l'hypothèse nulle est « $H_0 : p = p'$ », où p et p' sont les probabilités d'observer une naissance de garçon en cas de naissance légitime ou non. La mise en œuvre des tests impose au plan théorique de clarifier ce qu'on juge être le champ des possibles en dehors de H_0, qui définissent l'hypothèse alternative, notée H_1. Dans les deux exemples ci-dessus, on a ici choisi « $H_1 : p \neq 1/2$ » et « $H_1 : p \neq p'$ »[11].

[10] Une croyance dans cette région veut qu'un nourrisson risque le mauvais œil et la mort en étant exposé au regard des autres, ce qui est le cas lors de la séance de vaccination.

[11] D'autres champs des possibles plus restreints peuvent être considérés, tels « $H_1 : p>1/2$ » et « $H_1 : p>p'$ », mais nous n'en discuterons pas ici.

On sépare le champ des possibilités des résultats expérimentaux en deux lots ; l'un délimite, sous l'hypothèse nulle, la région de *rareté,* qui dans le langage standard des tests est appelée *région de rejet* ; le nombre α est appelé risque d'erreur ou risque d'erreur de première espèce. Le lot complémentaire, celui des valeurs que nous avons dites *communes,* est appelé région *d'acceptation*. Deux cas se présentent :

a) Les résultats sont dans la région de rejet

La convention de langage est de dire « l'hypothèse nulle est rejetée au risque α » ou suivant les cas, de parler de *différence significative au risque* α. Il y a cohérence sémantique entre *région de rejet* et la formulation « on rejette l'hypothèse nulle ».

Rejeter l'hypothèse nulle, c'est dire que l'hypothèse alternative est acceptée dans sa globalité. La logique de la manipulation des modèles est conforme à la logique usuelle ; pour le premier test, cela dit simplement que si la probabilité p ne peut pas valoir 1/2, alors elle est différente de 1/2.

b) Les données sont dans la région d'acceptation

Dans la situation précédente, on connaissait le risque de se tromper dans l'étalonnage du test avec un automate, risque quantifié par le nombre α. Ici c'est plus compliqué : si on fait produire des données (une liste de symboles F ou G) par un automate à partir d'une probabilité p' (inconnue) différente de 1/2 de produire G, la probabilité d'accepter l'hypothèse nulle (on dira en tel cas de se tromper) est une fonction de p'. Cette situation étant moins confortable que la précédente, on cherche à l'éviter.

De plus, la situation où les données sont dans la région d'acceptation n'implique pas qu'elles seraient dans la région de rejet de toute valeur du paramètre situé dans l'hypothèse alternative. En effet, pour tout jeu de données il existe plusieurs modèles avec qui elles sont compatibles (ainsi, les données utilisées sont compatibles à la fois avec $p = 105/205$ et avec $p' = 1051/2051$). Si on dit qu'on accepte l'hypothèse nulle, sans avoir clairement conscience qu'on ne teste pas une égalité au sens mathématique entre deux nombres, la logique usuelle est mise à mal : accepter H_0 ne veut pas dire qu'on rejette H_1, c'est-à-dire, l'égalité à 1/2 de la probabilité d'une naissance de garçon n'exclurait pas que cette probabilité soit différente de 1/2 ! Dans le langage standardisé des tests, pour sortir de cette apparence de contradiction, on s'interdit le plus souvent, notamment dans les ouvrages pédagogiques, de dire qu'on accepte l'hypothèse nulle. La formulation est « les données ne permettent pas de rejeter l'hypothèse nulle » et on insiste sur le fait que *ne pas rejeter* n'est pas synonyme d'*accepter* : cette distinction subtile est souvent ressentie comme de la sophistique, voire un parti pris de ne pas reconnaître les faits si ceux-ci ne vont pas dans le sens attendu. D'ailleurs pourquoi alors parler de région d'acceptation ?

En pratique, pour valider une idée, on choisit, lorsque c'est possible, l'hypothèse nulle de telle sorte que son rejet valide l'idée, cette validation étant entachée d'une incertitude qu'on maîtrise. Par exemple, pour prouver l'idée que le taux de masculinité diffère suivant qu'on a une naissance légitime ou illégitime, on prendra comme hypothèse nulle celle de l'égalité des taux. Le rejet de cette hypothèse nulle validera l'idée initiale. Pour tester l'efficacité d'un médicament, on prend comme hypothèse nulle qu'il est inefficace, de telle sorte que son rejet valide cette efficacité. Ce choix de l'hypothèse nulle est peu intuitif, mais est induit par la mathématique du test lui-même.

Par analogie, prenons un cas de diagnostic médical. En cas de suspicion d'une maladie, on va prendre comme hypothèse nulle un état de bonne santé (de non-maladie) : si les paramètres du patient sont en dehors des zones communes pour l'état de bonne santé, on rejette l'hypothèse nulle et on déclare le patient malade. Dans le cas où les paramètres du patient sont « normaux », imagine-t-on un médecin qui ne prendrait jamais la responsabilité de dire « avec l'information dont je dispose, je déclare le patient non malade » mais dirait systématiquement « l'information dont je dispose ne me permet pas de prouver la maladie» ? Sa responsabilité, dans le cadre d'éventuelles poursuites judiciaires, serait-elle moins engagée ? Pour le patient, les deux messages ne se valent pas du tout !

On trouvera des éléments historiques sur l'origine de ce langage dans le paragraphe portant sur « la vraie valeur d'une probabilité ».

5.2 Validation d'un modèle

C'est ici simple. Prenons le cas de la valeur 105/205 du taux de masculinité : s'il est compatible avec les données, on l'accepte et on l'utilise. Sinon, on le rejette, on en cherche un autre. Ici, on ne s'embarrasse plus de discours complexes et élégants. On avance résolument, sachant qu'un modèle peut être réfuté.

5.3 En pratique

En pratique, on a recours aux deux usages des tests dans un même mouvement. Ainsi, pour tester si les garçons et les filles ont en moyenne la même taille à la naissance, on est amené, pour pouvoir comparer ces moyennes[12], à commencer par comparer la dispersion des données (mesurée par un paramètre appelé écart-type). Dans un premier temps, on cherche à valider les modèles où les dispersions sont égales. On met en œuvre un test, et si les données sont compatibles avec cette égalité de dispersions, on accepte l'hypothèse nulle sans sourciller. Ensuite, on prend comme hypothèse nulle l'égalité des moyennes théoriques et là on devient précautionneux. On se limite, si c'est le cas, à l'énoncé « les données ne permettent pas le rejet de l'hypothèse nulle » et on évite de dire qu'on accepte celle-ci : cette différence de traitement langagier d'un test à l'autre n'en facilite pas la compréhension !

Ces deux usages des tests (confronter une idée, valider un modèle) ont été source de nombreux débats et conflits entre les pionniers de la première moitié du 19^{e} siècle (cf. Armatte, 2006). Pour R. A. Fischer, les tests (on parlait alors de test de significativité) ne trouvent leur emploi que pour rejeter l'hypothèse nulle. Quelques années plus tard d'autres auteurs, dont E. S. Pearson et J. Neyman, ont décidé de prendre en considération la situation de validation de modèles, d'où l'introduction d'un autre risque (hélas fonctionnel comme nous l'avons vu plus haut). Les tests de significativité sont alors devenus des « tests d'hypothèses ». La première réaction de R. A. Fisher est catégorique :

> « Neyman, thinking he was correcting and improving my own early work on tests of significance, as a mean of "improvement of natural knowledge", in fact reinterpreted them in terms of that technological and commercial apparatus which is known as an acceptance procedure. Now, acceptance procedures are of great importance in the modern world. When a large concern like the Royal Navy receives material from an engineering firm it is, I suppose, subject to sufficiently careful inspection and testing to reduce the frequency of the acceptance of faulty or defective consignments… But the logical differences between such an operation and the work of scientific discovery by physical or biological experimentation seem to me so

[12] On compare alors en fait des *moyennes théoriques*, appelées *espérances*, comme on a comparé des *chances théoriques* appelées *probabilités*.

wide that the analogy between them is not helpful, and the identification of the two sorts of operations is decidedly misleading. »

R. A. Fisher

Croyant amender et améliorer mes premiers travaux sur les tests de significativité employés pour « augmenter le savoir venu des faits », Neyman a de fait reformulé ceux-ci en termes d'outils techniques et à visée commerciale, connus sous le nom de procédures d'acceptation. Ces procédures sont d'une grande importance aujourd'hui, dans le monde moderne. Si une grande firme, comme la Royal Navy, réceptionne du matériel venant d'une entreprise industrielle, je conçois qu'il faille en faire une inspection assez soignée et faire des tests afin de pouvoir réduire la fréquence de livraisons contenant du matériel défectueux… Mais les différences logiques entre de telles opérations et les découvertes scientifiques en physique ou en biologie expérimentales me paraissent si énormes que les analogies entre elles n'apportent rien, et les identifier est carrément trompeur.

Aujourd'hui, le cadre scientifique est celui de la manipulation de modèles et non plus celui de l'axiome d'existence d'une *vraie valeur* d'une probabilité. La preuve statistique étant un élément de validation ou de rejet d'une hypothèse théorique, plus ou moins puissant suivant le nombre des données dont on dispose, les positions de Neyman et Fischer ne s'opposent plus. Il est temps de les unifier et de dire dans toutes les situations : « les données sont (resp. ne sont pas) compatibles avec l'hypothèse nulle », ou « au vu des données, on accepte (resp. on rejette) l'hypothèse nulle ».

La phase d'interprétation de la preuve statistique qui suit la mise en œuvre du test permet de situer la force ou la fragilité du résultat même du test[13]. Cette dernière phase est aussi le temps du retour éventuel sur les mesures recueillies. Dire que le résultat d'un test statistique quel qu'il soit est un *élément de preuve* et non une preuve définitive rend bien compte qu'on est dans le domaine de l'incertain. Mais cela sous-entend aussi un cadre idéal où les données numériques ne sont pas contestables : les définitions de ce qui est à mesurer ne sont pas ambigües, l'expérience est sans artefact, etc. Le consensus explicite de la priorité des faits mesurés sur la théorie ne saurait s'appliquer systématiquement, aveuglément. Les « grandes théories », et le *savoir solidifié* ne sont pas, dans un premier temps, *mis à la merci* des données expérimentales. Ainsi, en septembre 2011, des mesures issues de l'expérience internationale OPERA (voir référence) s'avéraient incompatible avec la théorie de la relativité restreinte dans laquelle la vitesse de la lumière ne pouvait pas être surpassée. La communauté scientifique a donné priorité à la théorie, la chaîne de production des données de l'expérience a été passée au crible. Et effectivement, une faille a été détectée, qui a permis de comprendre comment les résultats expérimentaux avaient été faussés.

5.4 D'où vient l'axiome de « vraie valeur » d'une probabilité ?

« For the logical fallacy of believing that a hypothesis has been proved to be true, merely because it is not contradicted by the available facts, has no more right to insinuate itself in statistical than in other kinds of scientific reasoning… It would therefore add greatly to the clarity with which the tests of significance are regarded if it were generally understood that tests of significance, when used accurately, are capable of rejecting or invalidating hypotheses, in so far as they are contradicted by the data: but that they are never capable of establishing them as certainly true. »

R. A. Fisher

Car l'erreur de logique consistant à croire qu'on a démontré qu'une hypothèse était vraie, simplement parce que les données dont on dispose ne la contredisent pas, n'a pas plus de raison de s'insinuer dans le raisonnement statistique que dans tout autre raisonnement scientifique… Cela clarifierait beaucoup la

[13] D'autres interprétations dites Bayésiennes attribuent des probabilités aux hypothèses testées, qui évoluent à mesure que sont recueillis des faits expérimentaux.

vision que l'on a des tests d'inférence si on comprenait qu'en en faisant un bon usage on est capable de rejeter ou d'invalider une hypothèse, dans la mesure où elle est contredite par les faits : mais les tests ne peuvent jamais établir avec certitude qu'une hypothèse est vraie.

Pour situer l'élaboration de la rhétorique en jeu dans le domaine des tests statistiques, plaçons-nous dans la première moitié du 20e siècle, à une époque où le recours à la modélisation était peu répandu (en dehors des grandes théories telles celles de la physique) et où les images mentales associées à la notion de probabilité étaient liées à des tirages de boules dans des urnes. Ainsi, pour concevoir ce qu'est la probabilité d'une naissance de garçon, on imaginait que *tout se passe comme si* on tirait avec remise une boule dans une urne dont une proportion p des boules était marquée d'une lettre G, et les autres d'une lettre F. La *vraie valeur* de la probabilité est alors cette proportion p. L'urne dont on parle est cependant fictive, la valeur correspondante de la probabilité l'est donc tout autant. Parler de la *vraie valeur* d'une probabilité, c'est un peu comme parler de la vraie valeur de la longueur de la corne de la licorne.

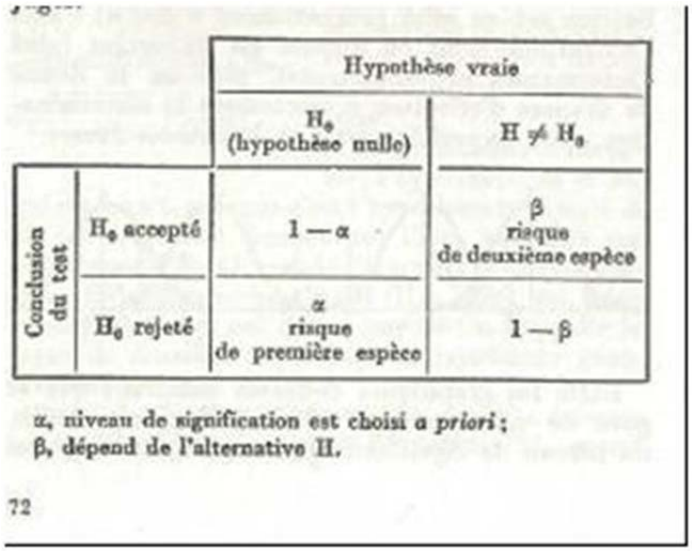
Parler de *vraies valeurs* au début du 20e siècle s'inscrivait dans la continuité du 19e siècle où la science s'était vue assigner la tâche d'estimer les constantes de la nature (Babagge, 1863) dont la *vraie probabilité* d'une naissance de garçon faisait partie. On a pensé à cette époque qu'il existait une *vraie* valeur des probabilités comme il *existe* une notion de *triangle en soi*, les triangles que l'on peut dessiner ou rencontrer dans la nature n'en étant que des images (des ombres de l'idée de triangle). Si on continue le parallèle avec la vision platonicienne, on peut dire que les concepts probabilistes sont les idées, les invariants, autour desquels fluctuent les quantités mesurées (fréquences, moyennes, etc.). La Statistique, interface de la théorie et des données d'observation, n'entre pas dans l'espace du paradigme platonicien. La démarche jusque-là impensée de quantification, celle des probabilités ou celle de la taille de l'homme *idéal* d'Adolphe Quetelet (Desrosières, 2002), introduite largement au 19e siècle avec la notion de *vraie valeur*, a cependant ouvert la voie au changement de paradigme que constitue la modélisation.

L'axiome d'existence d'une *vraie valeur* de chaque probabilité, et donc de la *véracité* des hypothèses testées, a paru de plus être doué de vertus pédagogiques fortes. Elle permet de parler de la probabilité β (appelé risque de deuxième espèce) de se tromper en acceptant à tort l'hypothèse testée, comme on peut le voir dans le tableau ci-dessous, reproduit dans de très nombreux ouvrages d'enseignement de la Statistique. La figure est extraite du livre « La Statistique », publié en 1947 par André Vessereau aux éditions des Presses Universitaires de France dans la collection « Que sais-je ». Ce petit livre, par ailleurs remarquable de clarté, a été l'objet de 21 éditions entre 1947 et 2012, et s'est vendu à plus de 180 000 exemplaires : il témoigne parfaitement de la rhétorique de la preuve en statistique dans la seconde moitié du 20e siècle.

Dans une optique de modélisation, ce qui correspond au nombre β devient, comme nous l'avons vu ci-dessus, une fonction, objet plus complexe à appréhender.

Le langage de la vraie valeur a eu cependant un prix élevé en termes de perte de sens de la preuve par les chiffres. Un exemple en est le suivant : pour comparer deux médicaments, il suffirait de comparer les *vraies valeurs* de leurs probabilités de guérison. Ces médicaments n'étant pas identiques, l'égalité stricte de ces *vraies probabilités* est une hypothèse en fait insensée ; de plus, la démontrer (i.e. démontrer l'égalité entre deux nombres, c'est-à-dire deux objets du monde mathématique) relèverait du seul champ de la théorie et non de celui de l'expérience. Avec l'axiome de la *vraie valeur*, l'hypothèse nulle testée est soit *fausse* (en fait

insensée), soit à jamais indémontrable par l'expérience. Ceci a eu un impact considérable au siècle dernier, dans les politiques de publications scientifiques, notamment en médecine. En effet, l'hypothèse d'égale efficacité des médicaments ne pouvant pas être *vraie,* si on ne la rejette pas, c'est finalement dû à un outil de preuve inefficace, et notamment à un nombre de données insuffisant. D'où l'émergence d'un *biais de publication* (*biais de tiroir* chez les sociologues) : une étude testant une question n'est publiée que si le résultat principal obtenu est prouvé en rejetant une hypothèse nulle (avec un risque α au plus égal à 0,05).



Conclusion du test	Hypothèse vraie : H_0 (hypothèse nulle)	Hypothèse vraie : $H \neq H_0$
H_0 accepté	$1 - \alpha$	β risque de deuxième espèce
H_0 rejeté	α risque de première espèce	$1 - \beta$

α, niveau de signification est choisi *a priori* ;
β, dépend de l'alternative H.

72

FIGURE 3 – *Un tableau classique des livres sur l'enseignement de la statistique*

6 En guise de conclusion

Si les métaphores des urnes restent pertinentes dans le chapitre des sondages et de certaines études de population, l'usage systématique de *populations fictives* et l'axiome de l'existence d'une « vraie valeur » n'ont pas lieu de perdurer. Faire évoluer le langage et les représentations n'est pas un désaveu des travaux antérieurs : élaborer une procédure de grande portée, la faire exister au plan mathématique et mettre en place un langage approprié ne se fait pas en un seul trait de pinceau.

Nous proposons donc de ne plus utiliser le vocabulaire de *rejet* ou de *non-rejet-ne-valant-pas-acceptation* de l'hypothèse nulle. Cette formulation, par l'absence de mention du jeu de données utilisé, donne d'ailleurs à la conclusion du test une teinte d'absolu qui en fausse le sens et la portée. En remplacement, on peut parler de données compatibles ou non avec l'hypothèse nulle. Les énoncés des conclusions possibles du test sont alors de même nature. Par contre, les deux risques associés, celui de conclure à tort à l'incompatibilité ou à la compatibilité ne le sont pas puisque le premier est un nombre et le second une fonction. Ces risques ne sont pas liés à la « vraie valeur » du paramètre d'un modèle, mais se réfèrent à une situation d'étalonnage du test où on fait produire (simuler) des données à un automate en contrôlant les paramètres de sa production.

La montée en puissance des moyens informatiques et la qualité des logiciels libres de statistique permet de choisir en toute occasion dans une batterie impressionnante de tests statistiques l'élément qui semble adapté. Les calculs, même longs et complexes, sont presque instantanés : nos *esclaves numériques* sont aussi rapides qu'infatigables. Par contre, la pensée

demande, aujourd'hui comme hier, du temps pour se déployer. Pendant des siècles, il a été plus rapide et plus aisé de penser que de calculer et aujourd'hui, c'est l'inverse. D'où une propension à être rudimentaire au niveau des phases de réflexion constitutives du processus de preuve statistique. Cette propension nourrit indirectement mais grassement les représentations de la Statistique comme art de faire dire ce qu'on veut aux chiffres. La Statistique n'est ni un art du mensonge, ni une pourvoyeuse de vérités universelles ou intemporelles. Reprenant les termes de F. Jacob (1997) ci-dessous, nous dirons qu'elle éclaire les *sciences de nuit* et laisse dans les *sciences de jour* (celles du savoir solidifié) une trace plus ou moins palpable.

> « La science de jour met en jeu les raisonnements qui s'articulent comme des engrenages, des résultats qui ont la force de la certitude. (…) La science de nuit, au contraire, erre à l'aveugle. Elle hésite, trébuche, recule, transpire, se réveille en sursaut. Doutant de tout, elle se cherche, s'interrompt, se reprend sans cesse. C'est une sorte d'atelier du possible où s'élabore ce qui deviendra le matériau de la science. »
>
> F. Jacob (1997), *La souris, la mouche et l'homme*

Références

[1] Arbuthnot, J. (1711), *An argument for Divine Providence, taken from the constant regularity observed in the births of both sexes*, From: *Philosophical Transactions of the Royal Society of London*, **27**, 186-190. Réédité dans Kendall, M. G. and R. L. Plackett (eds), *Studies in the History of Statistics and Probability Volume II* (1977), High Wycombe, Griffin, 30-34.

[2] Armatte, M. (2006), Contribution à l'histoire des tests laplaciens, *Mathematics and Social Sciences*, **176**(4), 117-133, http://www.ehess.fr/revue-msh/pdf/N176R1254.pdf

[3] Attané, I. (2008), *La Chine, un géant démographique aux pieds d'argile*, Fiche d'actualité de l'INED, http://www.ined.fr/fichier/t_telechargement/30988/telechargement_fichier_fr_fiche_actualit.2.3.pdf

[4] Babbage, C. (1863), *Sur les constantes de la nature, classe des mammifères*, http://www.sudoc.abes.fr/DB=2.1/SRCH?IKT=12&TRM=020558236

[5] Banerjee, A. et E. Duflo (2011), *Repenser la pauvreté*, Edition du seuil, Paris.

[6] Brian, E. et M. Jaisson (2007), *Le sexisme de la première heure*. *Hasard et sociologie*, Ed. Liber.

[7] CartoCrime.Net (2012), le portail géostatistique d'accès et de téléchargement des données localisées des crimes et délits enregistrés par la police et la gendarmerie nationales, http://www.cartocrime.net/Cartocrime2/index.jsf

[8] de Laplace, P. S. (1814), *Essai philosophique sur les probabilités*, Ed. Courcier. Ouvrage numérisé disponible en ligne.

[9] Desrosières, A. (1993), *La politique des grands nombres. Histoire de la raison statistique*, Édition La Découverte, Paris.

[10] Desrosières, A. (2002), Adolphe Quetelet, *Courrier des statistiques*, 104, INSEE, http://www.insee.fr/fr/ffc/docs_ffc/cs104a.pdf

[11] Desrosières, A. (2008a), *Pour une sociologie historique de la quantification. L'argument statistique I*, Presse de l'Ecole des mines, Paris.

[12] Desrosières, A. (2008b), *Gouverner par les nombres. L'argument statistique II*, Presse de l'Ecole des mines, Paris.

[13] Gavarret, J. (1840), *Principes généraux de statistique médicale – Développement des règles qui doivent présider à son emploi*. Ouvrage numérisé disponible en ligne.

[14] Hawlbachs, M. et A. Sauvy (1936), *Le point de vue du nombre.* Édition critique et commentée sous la direction de Brian, E. et M. Jaisson (2005), Classique de l'économie, INED, Paris.

[15] Haynes, J. D. et G. Rees (2006), Decoding mental states from brain activity in humans, *Nature Reviews*, *Neurosciences*, **7**, 523-534, http://www.utdallas.edu/~otoole/HCS6330_F09/17_Haynes_decoding_NNR_06.pdf

[16] Hemilä H., E. Chalker, B. Treacy, and B. Douglas (2009), *Vitamin C for preventing and treating the common cold*, Cochrane Database of Systematic Reviews, http://otorrinobrasilia.com/h1n1/cochrane_vitamin_c.pdf

[17] Hemilä, H. (2011), *Vitamin C and the common cold*. Bibliographie sur le site : http://www.mv.helsinki.fi/home/hemila/vitc_colds.htm

[18] Jacob, F. (1997), *La souris, la mouche et l'homme*, Éd. Odile Jacob, Paris.

[19] Les sceptiques du Québec. Site de l'association : http://www.sceptiques.qc.ca/dictionnaire/radin_dean.html

[20] Montesquieu, C. (1848), *De l'esprit des lois*. Ouvrage numérisé disponible en ligne.

[21] OPERA (2011), Communiqué de presse du CNRS sur l'expérience internationale OPERA, http://www2.cnrs.fr/presse/communique/2289.htm

[22] Pauling, L. (1971), The significance of the evidence about ascorbic acid and the common cold, *Proc. Nat. Acad. Sci. USA*, **18**(11), 2678-2681, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC389499/pdf/pnas00086-0061.pdf

[23] Soon, C. S., M. Brass, H. J. Heinze, and J. D. Haynes (2008), Uncounscious determinants of free decisions in the human brain, *Nature Neuroscience*, 11, 543-545, http://www.nature.com/neuro/journal/v11/n5/abs/nn.2112.html

[24] Süssmilch, P. (1741), *L'« ordre divin » aux origines de la démographie*. Traduit par M. Kriegel, présenté par J. Hecht, 3 volumes (1984), Classique de l'économie, INED, Paris.

[25] The Wall Street Journal (2012), *Murder in America. Interactive database*, http://projects.wsj.com/murderdata/#view=all

[26] Vessereau, A. (1947, 1re édition ; 2012, 21e édition), *La Statistique*, PUF, Collection Que sais-je ?, Paris.