Bayésianisme *versus* fréquentisme en inférence statistique

Jan Sprenger[*†]

Motivation du chapitre : la découverte de la particule de Higgs

Le bayésianisme et le fréquentisme, les deux grandes écoles d'inférence statistique, se distinguent par des postulats philosophiques et des méthodes mathématiques fondamentalement différentes. En un mot, alors que l'**inférence bayésienne** s'intéresse à la crédibilité d'une hypothèse, les fréquentistes se focalisent sur la fiabilité des procédures qui engendrent leurs conclusions. Plus précisément, une **inférence fréquentiste** est valide si, sur le long terme, la procédure qui en est à la base ne mène que rarement à des conclusions incorrectes. Par ailleurs, c'est cette manière de raisonner qui domine parmi les pratiques scientifiques. Pour comprendre le rôle du bayésianisme dans l'inférence statistique et évaluer sa capacité à améliorer le raisonnement scientifique, nous devons appréhender et apprécier les principes, les avantages et les inconvénients de l'école statistique fréquentiste. L'objectif de ce chapitre est de clarifier ces principes et de les comparer avec les principes de l'inférence bayésienne.

Pour comprendre cette différence entre bayésianisme et fréquentisme, il peut être utile de faire appel à la distinction, proposée par Royall[1], entre trois questions centrales :

1. Que devrions-nous *croire* ?
2. Que devrions-nous *faire* ?
3. Quand est-ce que les données comptent comme des *preuves* en faveur d'une hypothèse ?

En règle générale, aucune école statistique ne traite les trois questions sur un pied d'égalité. Les bayésiens se concentrent sur la première

---

1* Tilburg Center for Logic, General Ethics and Philosophy of Science (TiLPS), Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Courriel : j.sprenger@uvt.nl. Page web: www.laeuferpaar.de. † Texte traduit de l'anglais vers le français par Gauvain Leconte.

Richard Royall, *Statistical Evidence: A Likelihood Paradigm*, Londres, Chapman and Hall, 1997.

question – celle concernant la croyance rationnelle – car pour eux les hypothèses scientifiques font l'objet d'une incertitude subjective et personnelle. Ils cherchent à savoir comment les données devraient modifier notre degré de croyance en une hypothèse incertaine. C'est donc dans le cadre d'un modèle formel de croyance rationnelle, fourni par le calcul de probabilités, que les bayésiens répondent à la deuxième et à la troisième question – celles concernant les décisions rationnelles et les bonnes mesures d'évaluation de confirmation. De l'autre côté, les fréquentistes rejettent unanimement l'utilisation de degrés de croyance subjectifs en science. Pour eux la question : « que devrions-nous croire ? » n'est pas à strictement parler une question scientifique. Pourtant, ils sont toujours divisés quant à savoir laquelle des deux autres questions est la plus fondamentale. Les célèbres statisticiens Jery Neyman et Egon Pearson insistent par exemple sur le rôle de la théorie de la décision dans les inférences statistiques, et ont conçu des tests d'hypothèses scientifiques prenant la forme de procédures pour prendre des décisions fiables. D'autres statisticiens, comme Ronald Fisher, soulignent l'intérêt de l'évaluation des preuves (après expérience), et refusent de mêler la théorie de la décision aux inférences statistiques.

J'ouvre ce chapitre sur un exemple récent de l'impact des divergences entre bayésiens et fréquentistes sur l'évaluation des découvertes scientifiques. Le 4 juillet 2012, le CERN (Centre Européen pour la Recherche Nucléaire) à Genève, a créé la surprise en annonçant avoir découvert le *boson de Higgs* – une particule du modèle standard de la physique moderne recherchée depuis 1964. La découverte du boson de Higgs prouvant l'existence d'un mécanisme spécifique de brisure de symétrie électrofaible, elle fut d'une extrême importance pour la physique des particules.

Les chercheurs du CERN raisonnèrent en fréquentistes dans leur analyse statistique : en supposant l'inexistence du boson de Higgs, leurs résultats expérimentaux déviaient de plus de cinq écarts-types de l'espérance mathématique. Puisqu'un tel résultat n'a qu'une chance sur deux millions

de se produire par hasard, les statisticiens (et le service de presse) du CERN conclurent que le boson de Higgs avait bien été découvert.

Cette analyse suscita un vif débat entre statisticiens fréquentistes et bayésiens. Tony O'Hagan, statisticien bayésien réputé, envoya un e-mail à la liste de diffusion de l'*International Society for Bayesian Analysis* (ISBA) critiquant sévèrement l'intégralité de l'analyse statistique en question :

> Nous savons, d'un point de vue bayésien, que cela [une preuve fréquentiste standard, N.D.A.] ne fait sens que si (a) l'existence du boson de Higgs n'a qu'une très faible probabilité *a priori*, et/ou (b) les conséquences d'une annonce erronée de sa découverte seraient extrêmement dommageables. Aucune de ces deux conditions ne semble être réunie [...]. La communauté de la physique des particules tout entière a-t-elle épousé l'analyse fréquentiste ? Dans ce cas, quelqu'un a-t-il essayé d'expliquer la mauvaise science que cela représente ?[2]

Le message de O'Hagan déclencha de vifs échanges sur le forum de l'ISBA, auxquels prirent part des statisticiens et des physiciens des particules de premier plan. Le débat concernait d'abord un certain standard de preuve, mais, cette notion dépendant du cadre statistique choisi, il se transforma rapidement en une controverse généralisée sur les mérites des statistiques bayésiennes et fréquentistes. La découverte de la particule de Higgs illustre ainsi la manière dont l'interprétation d'un résultat scientifique fondamental dépend de questions méthodologiques concernant l'inférence statistique. On trouve de tels cas en dehors du domaine de la physique des particules : ils sont présents dans toutes les branches de la science où l'on utilise des méthodes statistiques, ce qui

---

2 Tony O'Hagan, Email sur les méthodes statistiques utilisées dans la découverte du boson de Higgs, posté *via* la liste de diffusion de l'International Society for Bayesian Analysis (ISBA), consulté sur www.isba.org le 6 janvier 2013.* NDT: le terme « preuve » traduit ici l'anglais « *evidence* », qu'il est toujours délicat de rendre en français car son sens change selon les contextes. « *Evidence* » désigne parfois les *données* qui confirment ou infirment une hypothèse scientifique, mais l'objet de ce chapitre est d'interpréter les procédures statistiques – bayésiennes ou fréquentistes – qui permettent d'évaluer si ces données rejettent ou non une hypothèse, c'est-à-dire la preuve que l'on tire de ces données.

inclut des problèmes aussi concrets que les procédures d'autorisation de mise sur le marché des médicaments.

Dans cette contribution, nous nous concentrons sur l'interprétation des preuves* statistiques. Il s'agit non seulement du champ le plus débattu entre bayésiens et fréquentistes, mais aussi du plus pertinent pour les statisticiens, les expérimentateurs, et les experts en politique scientifique. Ce chapitre est structuré comme suit : la section 1 récapitule les principes de l'inférence bayésienne. La section 2 explique les principes des procédures fréquentistes, alors que la section 3 porte sur la controverse entre bayésiens et fréquentistes. La section 4 traite du principe de vraisemblance et du principe de la règle d'arrêt, tandis que la section 5 passe en revue les objections contre le bayésianisme subjectifs, et les positions intermédiaires entre bayésianisme et fréquentisme.

1. L'inférence bayésienne

L'hypothèse de base de l'inférence bayésienne est que la probabilité doit être interprétée comme un degré de croyance rationnelle. Ainsi, le système de degrés de croyance d'un agent est représenté par une fonction de probabilité $p(\bullet)$, et $p(H)$ quantifie le degré auquel il croit que l'hypothèse H est vraie. On peut soutenir que ce degré de croyance doit se conformer au calcul des probabilités de différentes manières : soit en attirant l'attention sur les conséquences des actions guidées par des degrés de croyances qui ne s'y conforment pas (ce sont les théorèmes du pari hollandais[3]), soit en faisant remarquer qu'il en résulte une perte de précision[4].

---

3 *Cf.* John Kemeny, « Fair bets and inductive probabilities », *Journal of Symbolic Logic*, vol. 20 / 03, 1955, p. 263-273.

4 James Joyce, « Bayes' Theorem », in *Stanford Encyclopedia of Philosophy*, 2008. Hannes Leitgeb et Richard Pettigrew, « An objective justification of Bayesianism I: Measuring inaccuracy », *Philosophy of Science*, vol. 77 / 2, 2010, p. 201–235. Hannes Leitgeb et Richard Pettigrew, « An objective justification of Bayesianism II: The consequences of minimizing inaccuracy », *Philosophy of Science*, vol. 77 / 2, 2010, p. 236–272.

Les degrés de croyance probabilistes sont révisés à la lumière de nouvelles informations. Le degré de croyance en l'hypothèse *H* après avoir pris connaissance de la preuve *E* s'exprime au moyen de la probabilité conditionne de *H* sachant *E*, *p(H|E)* :

Conditionnement bayésien : le degré de croyance rationnelle en la proposition H après avoir appris ou découvert E est la probabilité conditionnelle de H sachant E : $p_{nouvelle}(H) = p(H|E)$.

p(H) et *p(H|E)*sont nommées **probabilités *a priori***et***a posteriori*** de *H*. Elles sont reliées par le **théorème de Bayes** :

$$p_{nouvelle} := p(H|E) = p(H)\frac{p(E|H)}{p(E)} \quad (1)$$

Les termes *p(E|H)*et *p(E|¬H)*sont nommés **vraisemblance** de *H* et ¬*H* relativement à E, c'est-à-dire la probabilité de la preuve observée *E* si l'on fait une hypothèse spécifique, ici *H* ou ¬*H*.

L'appellation« inférence bayésienne » est habituellement associée aux principes suivants :

- La représentation des degrés de croyance subjectifs par des probabilités.
- L'utilisation du conditionnement bayésien pour réviser rationnellement ses degrés de croyance.
- L'utilisation de la distribution de probabilité *a posteriori* pour évaluer les preuves, accepter les hypothèses et prendre des décisions.

Cependant, tous les bayésiens n'acceptent pas ces principes. Dans une conception bayésienne objective (par exemple celles proposées par Jeffreys[5] ou Bernado[6]), les probabilités *a priori* n'expriment pas l'incertitude subjective ; leur attribution peut aussi être guidée par des propriétés mathématiques commodes, comme l'invariance par transformation. Les bayésiens qui acceptent le principe d'entropie maximale refusent, eux, de considérer le conditionnement bayésien

5 Harold Jeffreys, *The theory of probability*, Oxford, Clarendon Press, 1939.

6 José M. Bernardo, « Integrated objective Bayesian estimation and hypothesis testing », *Bayesian statistics*, vol. 9, 2012, p. 1–68.

comme un guide pour déterminer les croyances rationnelles ; pour eux les degrés de croyance rationnelle devraient être les plus équivoques possible et se trouver autant que faire se peut dans la moyenne de ceux compatibles avec les preuves empiriques[7] (pour plus de détails sur ces approches, se reporter à la section 5). Enfin, certains statisticiens insistent plus sur les mesures bayésiennes du soutien d'une hypothèse (comme par exemple le facteur de Bayes[8]), que sur la valeur des distributions *a posteriori*. Pour simplifier les choses, nous nous focalisons ici sur la position subjectiviste classique en statistiques bayésiennes, qui est fondée sur la conjonction des trois principes ci-dessus (le lecteur peut se reporter à l'ouvrage de Howson et Urbach[9] pour une introduction philosophique à cette position, et à celui de Bernado et Smith[10] pour une approche plus mathématique).

2. Le fréquentisme : principes et tests de signification

Au XIX^e^ siècle, la théorie des probabilités a agrandi son domaine, partant des jeux de hasard pour s'appliquer progressivement aux problèmes d'analyse de données scientifiques, industrielles et administratives. Il s'agit peut-être de la date de naissance de la statistique inférentielle moderne. Elle apparut comme un outil pour gérer des ensembles de données complexes et pour quantifier le manque de précision des prédictions et des mesures. En raison de l'idéal dominant à cette époque, selon lequel la science doit aspirer à la certitude et fournir

7 Jon Williamson, *In Defence of Objective Bayesianism*, OUP Oxford, 2010.

8 *Cf.* Robert E. Kass et Adrian E. Raftery, « Bayes factors », *Journal of the american statistical association*, vol. 90 / 430, 1995, p. 773–795. Steven N. Goodman, « Toward evidence-based medical statistics. 2: The Bayes factor », *Annals of internal medicine*, vol. 130 / 12, 1999, p. 1005–1013.

9 Colin Howson et Peter Urbach, *Scientific Reasoning: The Bayesian Approach*, 3, La Salle, Open Court, 2006, 327 p.

10 José M. Bernardo et Adrian F. M. Smith, *Bayesian Theory*, Chichester, Wiley, 1994.

une vision objective et impartiale de la réalité, l'inférence bayésienne fut rejetée par de nombreux pères fondateurs de la statistique moderne. Les degrés subjectifs de croyance étaient, après tout, considérés comme très différents des preuves objectives. L'éminent statisticien Ronald Fisher parla même de « simples tendances psychologiques, les théorèmes qui les respectent n'ayant aucune utilité scientifique »[11]. Fisher expliqua qu'il ne croyait pas que le raisonnement bayésien fût logiquement invalide, mais qu'il était rare de disposer d'informations fiables sur lesquelles établir une distribution de probabilité *a priori* non arbitraire.

Mais comment réaliser des inférences allant des données à la théorie, si la route passant par le théorème de Bayes est fermée ? C'est ici qu'intervient la principale innovation des statistiques fréquentistes : la révision des croyances est remplacée par le **test d'hypothèse.** Dans leur article révolutionnaire de 1933, les statisticiens britanniques Jerzy Neyman et Egon Pearson lient l'inférence statistique à la prise de décision rationnelle, en développant une approche résolument fréquentiste des tests d'hypothèse : la décision d' « accepter » ou de « rejeter » une hypothèse devrait être prise de manière à minimiser la fréquence relative de mauvaises décisions dans une hypothétique série de répétitions du test. Autrement dit, la valeur d'une conclusion n'est pas assurée par sa forte probabilité *a posteriori*, mais par le fait qu'elle a été engendrée par une règle de décision fiable.

Prenons un exemple pour illustrer leur approche. Supposons que l'on ait à décider si un médicament passera la prochaine étape d'une procédure d'autorisation lourde et chère. Bien évidemment, on ne souhaite pas admettre un médicament qui ne soit pas supérieur aux traitements déjà existants en termes d'efficacité, d'effets secondaires, de coût, *etc*. D'un autre côté, on ne veut pas éliminer par erreur un médicament meilleur que ceux déjà existants. Ce sont les deux formes possibles d'erreur, communément appelées **erreur de type I** et **erreur de type II**.

La procédure standard pour de tels tests consiste à choisir une hypothèse par défaut ou **hypothèsenulle** H_0. Souvent cette hypothèse

11 Ronald A. Fisher, *The Design of Experiments*, Edinburgh, Oliver and Boyd, 1935, p. 6-7.

affirme que l'intervention expérimentale n'a aucun effet sur la variable cible. À l'inverse, l'**hypothèse alternative** affirme qu'un tel effet est présent. Alors qu'une erreur de type I correspond au rejet erroné de l'hypothèse nulle, l'erreur de type II consiste à l'accepter par erreur. Par convention, les taux d'erreur de type I acceptables sont fixés au **seuil** de 5%, 1%, ou 0,1%, bien que Neyman et Pearson insistent sur le fait que ces seuils n'ont aucune signification particulière, et que trouver le juste milieu entre les taux d'erreur de type I et de type II est une tâche qui dépend fortement du contexte dans lequel on se trouve. Il serait donc rationnel de se fier à cette procédure de test en raison de ses propriétés favorables sur le long terme :

> […] nous rejetterons *H* quand elle est vraie pas plus de, disons, une fois sur cent, et de plus nous pouvons avoir la preuve que l'on rejette *H* suffisamment souvent lorsqu'elle est fausse.[12]

Alors que Neyman et Pearson s'engagèrent dans le projet de trouver des tests aux propriétés optimales, un autre des aïeux de la statistique fréquentiste, l'éminent généticien et statisticien Ronald Fisher, s'opposa violemment à toute cette approche comportementale et décisionniste de l'inférence statistique. Pour Fisher, déterminer un taux d'erreur de type I acceptable implique une évaluation implicite de la gravité de cette erreur, imposant par là même, sur l'expérience scientifique en question, une évaluation de son utilité fondée sur ~~un barème utilitaire issu de~~ la théorie de la décision. Fisher soutint, au contraire, que

> Dans le champ de la recherche pure, aucune estimation du coût de conclusions erronées […] ne peut être plus qu'une vaine prétention, et, quoiqu'il en soit, une telle estimation serait inadmissible et sans aucune pertinence dans l'évaluation d'une preuve scientifique.[13]

Cette déclaration repose sur deux arguments. Le premier est que quantifier l'utilité globale d'une inférence statistique est une tâche quasi-

12 Jerzy Neyman et Egon Pearson, *Joint statistical papers*, Cambridge, Cambridge University Press, 1967, p. 142.

13 Ronald A. Fisher, *op. cit.*, p. 25-26.

impossible. Les conséquences ultimes de l'acceptation ou du rejet de l'hypothèse testée se situent au-delà de l'horizon épistémique des scientifiques. Les modèles de théorie de la décision ne sont donc pas aptes à décrire les inférences scientifiques. Le second argument est que les tests d'hypothèses statistiques devraient énoncer les *preuves* en faveur ou à l'encontre de l'hypothèse testée, et ne pas être obscurcies par la considération des conséquences pratiques découlant du fait de travailler avec telle hypothèse particulière. En d'autres termes, si les tests de Neyman-Pearson peuvent être utiles pour les contrôles de qualité industriels, ou dans d'autres contextes appliqués, ils n'examinent pas la *vérité* des hypothèses scientifiques, et ne sont pas des outils appropriés pour la recherche scientifique (pure).

Comme alternative, Fisher[14] inventa le paradigme des **tests de signification statistique**. Pour lui, l'objectif de l'analyse statistique consistait à évaluer la relation d'une hypothèse (nulle) à un corps de données observables. Cette hypothèse affirme habituellement l'absence d'un phénomène intéressant, par exemple l'absence de relation causale entre deux variables, de différence observable entre deux traitements, *etc*. En accord remarquable avec la méthodologie falsificationniste de Popper, Fisher affirme que le seul objectif d'une expérience est de « donner une chance aux faits d'infirmer une hypothèse nulle »[15] et que l'échec du rejet d'une hypothèse ne permet pas de conclure à une preuve positive en faveur de l'hypothèse (nulle) testée. Mais à la différence de Popper[16], Fisher vise la *démonstration* expérimentale et statistique de phénomènes. Il a donc besoin d'un critère pour distinguer les effets réels des artefacts expérimentaux. Le critère que suggère Fisher est l'incompatibilité de l'hypothèse nulle avec les données observées, incompatibilité mesurée par l'improbabilité des données sous l'hypothèse nulle :

14 Ronald A. Fisher, *Statistical methods and scientific inference*, New York, Hafner Press, 1956.

15 Ronald A. Fisher, *Statistical methods for research workers*, Edinburgh, Oliver and Boyd, 1925, p. 16.

16 Karl R. Popper, *Logik der Forschung*, Berlin, Akademie Verlag, 1934.

> Soit une possibilité exceptionnellement rare a eu lieu, soit la théorie [l'hypothèse nulle] n'est pas vraie.[17]

Ce schéma d'inférence élémentaire, que Hacking[18] nomme **la disjonction de Fisher**, est au centre des tests de signification statistique. Une possibilité si exceptionnellement rare a des conséquences à la fois épistémologiques et pratiques : en premier lieu, elle rend l'hypothèse nulle « objectivement sans crédibilité »[19], et, en second lieu, elle implique que l'hypothèse nulle devrait être traitée comme fausse. Il faut remarquer que l'approche de Fisher est essentiellement asymétrique : alors que le « rejet » discrédite fortement l'hypothèse nulle, l' « acceptation » signifie seulement que les faits ont échoué à réfuter l'hypothèse nulle. À l'inverse, l'interprétation du résultat des tests de Neyman-Pearson est essentiellement symétrique. Ils attribuent un rôle explicite à l'hypothèse alternative là où l'approche de Fisher se focalise uniquement sur l'hypothèse nulle.

Le schéma d'inférence inhérent à la disjonction de Fisher, une sorte de *modus tollens* probabiliste[20], a été fréquemment critiqué. Hacking[21] a souligné les problèmes inévitables auxquels on est confronté lorsqu'on cherche à clarifier l'expression « exceptionnellement rare ». Si elle signifie que l'événement observé doit être exceptionnellement peu probable comparé aux autres événements, certaines hypothèses statistiques ne pourraient jamais être testées. La distribution uniforme sur un ensemble fini d'événements, par exemple, attribue une probabilité égale à toutes les observations. Comment pourrions-nous alors tester – et peut-être rejeter – une telle hypothèse dans la conception de Fisher ?

---

17 Ronald A. Fisher, *op. cit.*, p. 39.

18 Ian Hacking, *Logic of Statistical Inference*, Cambridge, Cambridge University Press, 1965, 244 p.

19 Stephen Spielman, « The logic of tests of significance », *Philosophy of Science*, 1974, p. 211–226, p. 214.

20 Voir également : Donald A. Gillies, « A falsifying rule for probability statements », *British Journal for the Philosophy of Science*, 1971, p. 231–261.

21 Ian Hacking, *op. cit.*, p. 81-82.

Pour approfondir cette remarque, imaginons que l'on teste l'hypothèse selon laquelle les résultats des lancers d'une pièce de monnaie spécifique sont indépendants et identiquement distribués, avec une égale probabilité de pile et de face. Comparons maintenant deux séries de résultats : 'FPPFPPPFFF' et 'PPPPPPPPPP'. La probabilité des deux événements sous l'hypothèse nulle est la même : $(1/2)^{10} = 1/1024$. Pourtant, la seconde série, et non la première, semble plaider contre l'hypothèse nulle. Pourquoi est-ce le cas ? Parce qu'implicitement nous avons précisé *la manière dont les données sont exceptionnelles* : nous nous intéressons à la propension θ de la pièce à tomber sur pile plutôt qu'à l'indépendance entre les lancers ou tout autre présupposé sur lequel repose le test. On peut donc restreindre notre attention au nombre de lancers donnant face (F), et en effet $\{F=0\}$ est bien moins probable que $\{F=5\}$[22].

Il semble que l'on ne puisse appliquer les tests de signification sans introduire implicitement des hypothèses alternatives ; ici, par exemple, que la pièce est biaisée vers pile ou vers face. Spielman[23] prolonge cet argument dans une vaste analyse logique des tests de signification : inférer d'un résultat peu probable la présence d'un effet significatif *présuppose* que le résultat observé est bien plus probable sous une hypothèse alternative que sous l'hypothèse nulle. Et en effet, les approches fréquentistes modernes, comme la théorie statistique de l'erreur proposée par Mayo[24], prennent cela explicitement en compte en concevant l'inférence statistique de manière contrastée, en ce qu'un test concerne toujours un écart avec l'hypothèse testée. Il est important de garder cela à l'esprit lorsque l'on compare les mesures de preuve fréquentistes et bayésiennes.

---

22 Richard Royall, *op. cit.*, chap. 3.

23 Stephen Spielman, *op. cit.*

24 Deborah G. Mayo, *Error and the growth of experimental knowledge*, Chicago, University of Chicago Press, 1996.

3. Le fréquentisme : les valeurs-p

Les tests de signification statistiques dans la tradition de Fisher sont sans doute l'outil méthodologique le plus populaire dans les pratiques statistiques. Il y a cependant d'importantes différences entre la conception originelle de Fisher, discutée ci-dessus, et la pratique actuelle des tests de signification en science, qui est une forme hybride entre les tests de l'école de Fisher et de l'école de Neyman-Pearson, et dans laquelle le concept de ***valeur-p*** joue un rôle central.

Pour expliquer ces différences, il faut distinguer entre **problèmes de test unilatéral** et **bilatéral**. Un problème unilatéral consiste à demander si un paramètre inconnu est plus ou moins grand qu'une valeur particulière ($\theta \leq \theta_0 versus \theta > \theta_0$), alors qu'un problème bilatéral (ou test d'hypothèse nulle ponctuelle) consiste à demander si un paramètre θ est exactement égal à θ_0 : $H_0 : \theta = \theta_0 versus\ H : \theta \neq \theta_0$. Le test bilatéral peut être utilisé pour poser différentes questions. En premier lieu, la question de savoir s'il y a ou non un « effet » dans les données (par exemple, si l'hypothèse nulle désigne l'absence de relation causale). Ensuite, la question de savoir si H_0 est un substitut approprié à $H_0 \vee H_1$, c'est-à-dire si l'hypothèse nulle est une idéalisation, dont les prédictions sont suffisamment précises, d'un modèle statistique plus général.

Illustrons le concept central des tests de signification modernes – la *valeur-p* – par un problème de test bilatéral. A nouveau, on veut réaliser l'inférence à la présence d'un effet significatif si la divergence entre les données $x \coloneqq (x_1, \ldots, x_N)$, correspondant aux N réalisations d'une expérience, et l'hypothèse nulle $H_0 : \theta = \theta_0$ est suffisamment grande. Supposons que la variance σ^2 de la population est connue. On mesure alors la divergence des données x par rapport à la valeur moyenne postulée $\theta_0$au moyen de la statistique de test standardisée :

$$z(x) = \frac{\frac{1}{N}\sum_{i=1}^{N} x_i - \theta_0}{\sqrt{N.\sigma^2}} \qquad (2)$$

On peut réinterpréter l'équation (2) ainsi :

$$z = \frac{effet\ observé - effet\ supposé}{écart - type} \qquad (3)$$

Déterminer si un résultat est significatif ou non dépend de la *probabilité critique* dite *valeur-p*, ou **seuil de signification observé**, c'est-à-dire la probabilité, si l'on fait l'hypothèse nulle, d'observer un écart au moins aussi extrême que celui qui est réellement observé. Cette valeur dépend de *z* et peut être calculée ainsi :

$$p := p(|z(X)| \geq |z(x)|) \tag{4}$$

La figure 1 représente un seuil de signification observé *p = 0.072* sous la forme d'une intégrale de la fonction de distribution de probabilité. Evidemment, plus la valeur p est faible, plus la présomption contre l'hypothèse nulle est forte.

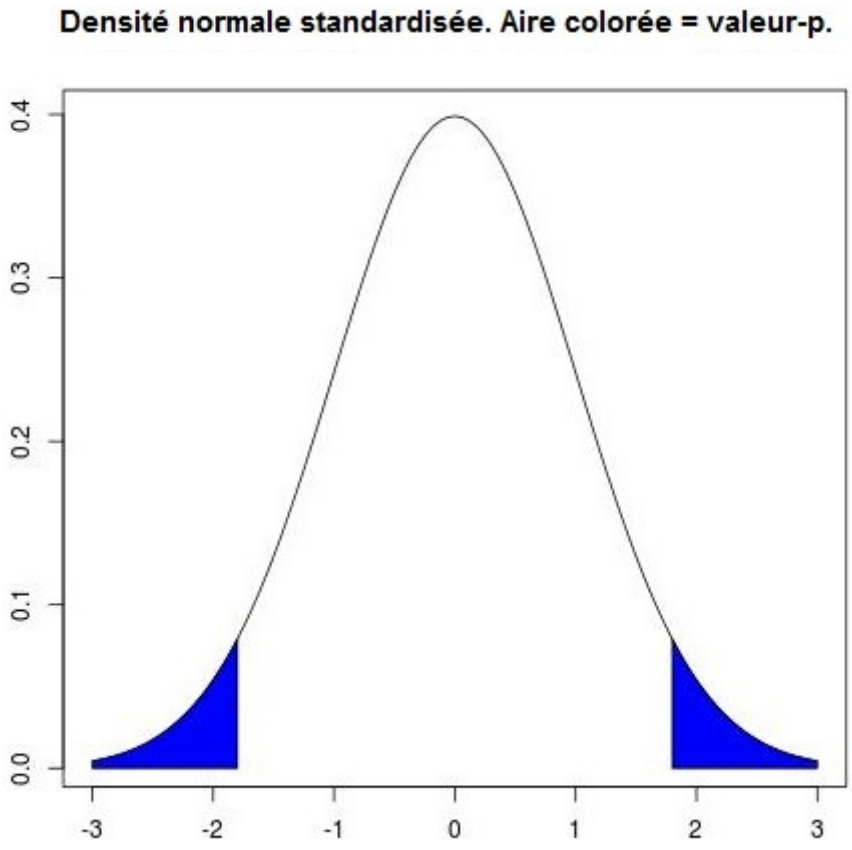

Figure 1 : la fonction de densité de probabilité de l'hypothèse nulle $H_0 : X \sim N(0, 1)$, testée contre l'hypothèse alternative $H_1 : X \sim N(\theta, 1)$. L'aire colorée représente la valeur p calculée pour la donnée observée $x = 1.8$ $(p=0.072)$.

Pour le praticien fréquentiste, les valeurs-p sont des mesures pratiques, reproductibles et objectives des preuves contre l'hypothèse nulle : elles peuvent être calculées automatiquement une fois que le modèle statistique est spécifié, et ne dépendent que de la distribution d'échantillonnage des données sous l'hypothèse H_0.Fisher les interprétait comme « une mesure des motivations rationnelles pour l'augmentation du *doute* [envers l'hypothèse nulle] qu'elles provoquent »[25].

Les vices et les vertus des tests de signification et des valeurs-p ont été examinés en détail dans la littérature sur ce sujet et on ne peut accorder ici la place nécessaire à un débat exhaustif (le lecteur peut se reporter à ce sujet aux travaux de Cohen[26] et de Harlow[27]). Cependant, certaines critiques importantes, ainsi que la relation de ces tests à l'inférence bayésienne, sont discutées ci-dessous.

3.1 L'erreur de la négligence du taux de base

Le principal problème des valeurs-p est sans doute d'ordre pratique : de nombreux chercheurs ne parviennent pas à les interpréter correctement. Bien souvent, une faible valeur-p (par exemple $p < 0.001$) est comprise comme l'affirmation que l'hypothèse nulle a une probabilité *a posteriori* inférieure à ce nombre[28]. Mais la valeur-p est essentiellement la probabilité conditionnelle d'un phénomène E si l'hypothèse nulle est donnée, $p(E|H_0)$, ce qui est très différent de la probabilité conditionnelle de l'hypothèse nulle si le phénomène est donné, $p(H_0|E)$.

---

25 Ronald A. Fisher, *op. cit.*, p. 43.

26 Jacob Cohen, « The earth is round (p<. 05). », *American psychologist*, vol. 49 / 12, 1994, p. 997-1001.

27 Lisa L. Harlow, Stanley A. Mulaik et James H. Steiger, *What If There Were No Significance Tests?*, Mahwah/NJ, Erlbaum, 1997, 472 p.

28 Voir par exemple à ce sujet : Michael W. Oakes, *Statistical inference: a commentary for the social and behavioural sciences*, New York, Wiley, 1986, 208 p.. Fiona Fidler, *From Statistical Significance to Effect Estimation.*, University of Melbourne, 2005.

Illustrons cette erreur de raisonnement, et ce qui la rend si fréquente, par un exemple simple. Considérons un don de sang, soumis à un test systématique de dépistage du VIH. Soit l'hypothèse nulle, qui énonce que le donneur n'a pas contracté le VIH. Les résultats du test sont corrects dans 99% des cas, qu'il soit infecté par le VIH ou non. Supposons que le test aboutisse à un résultat positif. Cela constitue, si l'on fait l'hypothèse nulle, une possibilité extrêmement rare, tandis que si l'on fait l'hypothèse alternative, elle est très probable. Le donneur devrait-il être convaincu qu'il a contracté le VIH, si la prévalence du VIH dans la population est de 0,01% ?

Un calcul bayésien aboutit à la conclusion, peut-être surprenante, qu'il devrait être assez certain de *ne pas avoir* contracté le VIH :

$$p(\textit{contraction du VIH}|\textit{test positif}) = \left(1 + \frac{p(\textit{test positif}|\textit{pas d'infection})}{p(\textit{test positif}|\textit{contraction du VIH})}\frac{p(\textit{pas d'infection})}{p(\textit{contraction du VIH})}\right)^{-1}$$

$$= \left(1 + \frac{0{,}01}{0{,}99}\frac{0{,}9999}{0{,}0001}\right)^{-1} \approx 0.01$$

En d'autres termes, la preuve en faveur de la contraction du VIH est plus que compensée par le très faible taux de base d'infection dans la population en question. Rejeter directement l'hypothèse nulle sur la base d'un résultat « significatif » n'est donc pas une inférence probabiliste valide, même si le résultat est probable sous l'hypothèse alternative. Puisque cette erreur de raisonnement est causée par la négligence du taux de base dans les populations, elle est connue sous le nom d'**erreur de négligence du taux de base**[29].

Malgré de persistants efforts pour éliminer cette erreur de négligence du taux de base, elle demeure monnaie courante parmi les praticiens. Certains spécialistes ont soutenu qu'il s'agissait d'un effet du caractère contre-intuitif de la conception fréquentiste dans son entier. Le psychologue allemand Gerd Gigerenzer[30], par exemple, soutient que les

29 Steven N. Goodman, « Toward evidence-based medical statistics. 1: The P value fallacy », *Annals of internal medicine*, vol. 130 / 12, 1999, p. 995–1004.

30 Gerd Gigerenzer, « The bounded rationality of probabilistic mental models », in K.I. Manktelow, D.E. Over. *Rationality: psychological and philosophical perspectives*, Florence, KY, US, Taylor & Frances/Routledge, 1993.

scientifiques sont avant tout intéressés par le caractère crédible et soutenable d'une hypothèse, et non la probabilité des données sous l'hypothèse nulle. La question est alors la suivante : comme devrions-nous relier les valeurs-paux mesures bayésiennes de la présomption en faveur d'une hypothèse ? Après tout, les analyses bayésiennes et fréquentistes devraient s'accorder lorsque les distributions de probabilité *a priori* peuvent être objectivement établies. La comparaison entre les mesures de soutien bayésiennes et fréquentistes est donc une entreprise qui ne se restreint pas aux mathématiques, mais qui a aussi une importance philosophique.

3.2 Valeur-p et mesures de soutien bayésiennes

Les bayésiens fondent leurs croyances et leurs décisions sur la distribution de probabilité *a posteriori* des hypothèses en jeu ; mais qu'utilisent-ils comme mesure de preuves afin d'exprimer combien les données plaident pour l'hypothèse nulle par rapport à l'hypothèse alternative (ou vice versa) ? Une mesure spécifique est utilisée de manière quasi-universelle en statistiques bayésiennes : le **facteur de Bayes**, c'est-à-dire le rapport des probabilités *a priori* et *a posteriori* entre l'hypothèse nulle $H_0 : \theta \in \Theta_0$ et l'hypothèse alternative $H_1 : \theta \in \Theta_1$, sachant la donnée x[31].

$$B_{01}(x) := \frac{p(H_0|x)}{p(H_1|x)} \cdot \frac{p(H_1)}{p(H_0)} = \frac{\int_{\theta \in \Theta_0} p(x|\theta)p(\theta)\,d\theta}{\int_{\theta \in \Theta_1} p(x|\theta)p(\theta)\,d\theta} \quad (5)$$

Ainsi, pour deux hypothèses composites H_0et H_1, le facteur de Bayes peut s'écrire comme le rapport des intégrales des vraisemblances, pondérées par la probabilité *a priori* de chaque hypothèse individuelle. Cette mesure est attirante pour plusieurs raisons. La plus cruciale est que l'on peut dériver la probabilité *a posteriori* d'une hypothèse H si l'on connait sa probabilité *a priorip(H)* et le facteur de Bayes de H par rapport à $\neg H$. Dans le cas d'hypothèses ponctuelles $H_0 : \theta = \theta_0$*versus* $H_1 : \theta = \theta_1$, le facteur de Bayes se ramène au **rapport de vraisemblances***L(x, H_0, H_1)*

31 Robert E. Kass et Adrian E. Raftery, *op. cit.*

= $p(x|H_0)/p(x|H_1)$, qui possède d'intéressantes propriétés d'optimalité en tant que mesure de preuve[32].

Il faut remarquer que les facteurs de Bayes et les valeurs-p peuvent substantiellement diverger. Pour illustrer ce paradoxe, prenons un exemple issu des recherches en parapsychologie[33]. Le cas en question implique le test de la prétention d'un sujet à pouvoir affecter une série de zéro et de un engendrée aléatoirement ($\theta_0 = 0,5$) grâce à ses pouvoirs extrasensoriels (PES). Le sujet affirmait pouvoir, par le seul usage de l'esprit, pouvoir faire dévier significativement la moyenne de l'échantillon de 0,5.

De très nombreuses données furent récoltées (N = 104 490 000) afin de tester cette hypothèse. La séquence de zéro et de un, $X_1, \ldots, X_N$, était décrite par un modèle binomial $B(\theta, N)$. L'hypothèse nulle affirmait que les résultats sont engendrés par une machine fonctionnant aléatoirement avec une probabilité $H_0 : \theta = \theta_0 = 1/2$, tandis que l'hypothèse alternative consistait en l'hypothèse non précisée $H_1 : \theta \neq 1/2$.

Jahn, Dunne et Nelson[34] rapportent que sur 104 490000 épreuves, 52 263471 un et 52 226 529 zéro furent observés. Un fréquentiste devrait alors calculer la valeur de z, qui est

$$z(x) = \sqrt{\frac{N}{\theta_0(1-\theta_0)}}\left(\frac{1}{N}\sum_{i=1}^{N} x_i - \theta_0\right) \approx 3,61$$

Il devrait donc rejeter l'hypothèse nulle, car cela produit une valeur-p très faible :

$$p := p(|z(X)| \geq |z(x)|) \ll 0,01$$

Ainsi, les données devraient interprétées comme une preuve convaincante en faveur de la présence de capacités extrasensorielles.

---

32 Richard Royall, *op. cit.*. Subhash Lele, « Evidence Functions and the Optimality of the Law of Likelihood », *The nature of scientific evidence: Statistical, philosophical, and empirical considerations*, 2004, p. 191–216.

33 R. G. Jahn, B. J. Dunne et R. D. Nelson, « Engineering anomalies research », *Journal of Scientific Exploration*, vol. 1 / 1, 1987, p. 21-50.

34 *Ibidem*.

Comparons à présent ce résultat à celui d'une analyse bayésienne. Jefferys[35] attribue à l'hypothèse nulle la probabilité conventionnelle positive $p(H_0) = \epsilon > 0$, une distribution *a priori* uniforme à l'hypothèse alternative, et calcule le poids de la preuve que x fournit à H_0 par rapport à H_1, qui est quantifié par le facteur de Bayes B_{01} :

$$B_{01}(x) := \frac{p(H_0|x)}{p(H_1|x)} . \frac{p(H_1)}{p(H_0)} \approx 12$$

Il s'ensuit que les données favorisent clairement l'hypothèse nulle par rapport à l'hypothèse alternative, et ne fournissent aucune preuve de la présence de PES.

Le fait que les inférences bayésiennes et fréquentistes puissent complètement diverger lorsque la taille de l'échantillon augmente, est un phénomène connu sous le nom de **paradoxe de Lindley[36]**. Comment l'expliquer ? Un facteur important est que les résultats statistiquement significatifs ne sont pas systématiquement de bons indicateurs de la **taille d'un effet**. Le générateur aléatoire automatique qui produit les séquences de un et de zéro n'est pas parfait en pratique, et souffre de légers biais. Un test de signification détectera ces biais dans un grand échantillon et conclura, avec un haut degré de confiance, qu'il faut rejeter l'hypothèse nulle H_0 selon laquelle la moyenne de la population est précisément de 0,5. Mais une telle conclusion estompe la différence entre signification statistique et signification scientifique : l'effet peut être négligeable. Par rapport à l'ensemble de toutes les hypothèses alternatives (dont certaines incluent des effets de taille non triviaux), H_0peut toujours être maintenue. C'est cette intuition qui nourrit l'analyse par facteur de Bayes, et qui explique la divergence entre les deux

35 William H. Jefferys, « Bayesian analysis of random event generator data », *Journal of Scientific Exploration*, vol. 4 / 2, 1990, p. 153–169.

36 Dennis V. Lindley, « A statistical paradox », *Biometrika*, 1957, p. 187–192.

résultats[37]. Pour résumer, une faible valeur-p n'indique pas, d'un point de vue bayésien, que l'hypothèse est moins tenable qu'auparavant.

Ce phénomène n'est pas un problème purement théorique : il mène fréquemment à l'appréhension erronée d'une signification statistique comme signalant un effet substantiel, ce qui met en péril la validité des conclusions qui en sont tirées[38]. En examinant minutieusement les pratiques statistiques de l'éminente revue scientifique d'économie *American Economic Review*, et en enquêtant sur les opinions des économistes à propos du sens de la signification statistique, McCloskey et Ziliak sont arrivés à la conclusion que la plupart des économistes ne connaissent pas la signification correcte des concepts statistiques. Concrètement, c'est la « pratique de l'astérisque » qui prévaut : dans les tables de corrélation par exemple, les résultats les plus significatifs sont marqués d'un astérisque, et ces résultats sont ceux qui sont supposés être réels, importants, et avoir une forte incidence sur la pratique. Mais un effet important et remarquable n'est pas forcément statistiquement significatif, de même qu'un effet statistiquement significatif peut être faible et peu intéressant (comme dans le cas de la PES).

Des recherches théoriques sont aussi menées qui visent à relier valeurs-p et probabilités *a posteriori*. Il s'avère que cela est souvent possible dans le cas des problèmes de tests unilatéraux[39], alors qu'habituellement elles divergent dans le cas des problèmes de test bilatéraux. Plus précisément, Berger et Sellke[40] ont montré que la valeur critique p est proportionnelle à la borne inférieure de la probabilité *a posteriori* de l'hypothèse nulle,

---

37 Jan Sprenger, « Testing a Precise Null Hypothesis: The Case of Lindley's Paradox », *Philosophy of Science*, vol. 80 / 5, 2014, p. 733–744. Christian Robert, « On the Jeffreys-Lindley Paradox », *Philosophy of Science*, vol. 81 / 2, 2014, p. 216–232.

38 Deirdre N. McCloskey et Stephen T. Ziliak, « The standard error of regressions », *Journal of Economic Literature*, 1996, p. 97–114. Stephen T. Ziliak et Deirdre N. McCloskey, *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, University of Michigan Press, 2008, 354 p.

39 George Casella et Roger L. Berger, « Reconciling Bayesian and frequentist evidence in the one-sided testing problem », *Journal of the American Statistical Association*, vol. 82 / 397, 1987, p. 106–111.

conduisant ainsi à une surestimation systématique des preuves contre cette hypothèse. Ce résultat semble suggérer une incompatibilité majeure entre mesures fréquentistes et bayésiennes de la présomption en faveur d'une hypothèse dans le cas de tests bilatéraux (pour plus de détails sur cette question, se reporter au chapitre de Christian Robert dans ce volume).

3.3 La plupart des résultats publiés sont-ils faux ?

Un des problèmes méthodologiques de l'usage des valeurs p, dû au fait qu'elles prennent leurs racines dans les tests de signification de Fisher, est que les résultats insignifiants (dont la valeur p est supérieure à 0,05) n'ont quasiment aucune chance d'être publiés. Ce problème est préoccupant sous (au moins) deux aspects : premièrement, même un résultat statistiquement insignifiant peut dissimuler un effet important et scientifiquement pertinent ; secondement, cela empêche l'appréciation de la présomption *en faveur de l'hypothèse nulle*. De précieuses ressources sont ainsi gaspillées, car différentes équipes de recherches, n'étant pas informées des efforts les unes des autres, reproduisent ces résultats non significatifs de nombreuses fois. De plus, la conception fréquentiste ne propose aucune formalisation logique d'inférence concluant d'un résultat non significatif à l'hypothèse nulle, si ce n'est qu'il échoue à la rejeter.

Cette asymétrie de l'inférence fréquentiste est au cœur de la célèbre thèse de John Ioannidis[41] selon laquelle « la plupart des résultats publiés sont faux ». Le raisonnement de Ioannidis est que de nombreuses hypothèses fausses peuvent être soutenues de manière erronée et offrir un résultat publiable. Si l'on cherche à tester les relations causales significatives dans un vaste ensemble de variables, la probabilité d'obtenir un faux positif est, pour des taux d'erreur de type I et II α et β,

40 James O. Berger et Thomas Sellke, « Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence », *Journal of the American Statistical Association*, vol. 82 / 397, mars 1987, p. 112-139.

41 John Ioannidis, « Why most published research findings are false », *PLoS medicine*, vol. 2 / 8, 2005, p. e124.

normalement plus grande que la probabilité d'obtenir une hypothèse vraie. En particulier, si *R* représente le rapport entre les relations vraies et fausses qui sont testées dans le cadre d'une recherche scientifique, *T* une relation causale particulière et *E* une preuve significative en faveur de *T*, alors

$$p(T|E) = \frac{p(E|T)p(T)}{p(E|T)p(T)+p(E|\neg T)p(\neg T)} = \frac{(1-\beta)R}{(1-\beta)\,R+\alpha} \quad (6)$$

Cette quantité est plus petite que *1/2* si et seulement si $R < \alpha/(1 - \beta)$, condition typiquement satisfaite si le seuil de standard de publication de résultats est fixé à $\alpha = 0{,}05$, et que la plupart des relations causales examinées ne sont pas substantielles. Ainsi la plupart des résultats publiés sont en réalité des artefacts issus des données manifestement faux. Cet effet est accru par de mauvaises pratiques de recherche qui mènent à des biais dans la sélection, le traitement et l'analyse des ensembles de données[42]. Il faut remarquer que ce phénomène n'est pas une caractéristique de l'activité scientifique en général, mais une spécificité due à la logique fréquentiste de l'inférence statistique : l'obtention d'un résultat significatif n'est tout simplement pas un très bon indicateur de la crédibilité d'une hypothèse. En réalité, les chercheurs ont souvent du mal à reproduire les résultats d'autres équipes, et les alternances entre enthousiasme et déception sont monnaie courante dans les domaines à la pointe de la recherche scientifique. Pour empêcher cela, l'utilisation de mesures bayésiennes de preuves peut être un remède approprié, car elles réalisent naturellement un compromis entre anticipations théoriques et résultats observés[43].

3.4 Les intervalles de confiance

Cette section se clôt par un mot sur les intervalles de confiance, fonctions qui ont pour argument une valeur observée x_0, et pour valeur un intervalle Cx_0. Ils sont souvent recommandés pour améliorer les tests

42 Gregory Francis, « The frequency of excess success for articles in Psychological Science », *Psychonomic bulletin & review*, 2014, p. 1–8.

43 Steven N. Goodman, *op. cit.*

d'hypothèse et pour résoudre les problèmes fondamentaux de la conception fréquentiste[44].

Un intervalle de confiance d'un niveau $\alpha \in [0, 1]$ne signifie pas que si l'on observe x_0, le paramètre θ se trouve dans l'intervalle Cx_0 avec la probabilité α. Après tout, les fréquentistes n'assignent aucune probabilité *a posteriori* à des valeurs spécifiques du paramètre étudié. Le niveau de confiance de l'intervalle indique plutôt quelque chose concernant la procédure utilisée pour construire cet intervalle : pour chaque θ, on construit un intervalle C_θ tel que les données x seront, sur le long terme, *cohérente* avec θ dans $100.\alpha$ % des cas. Si l'on projette l'ensemble $\{(x|\theta)| x \in C_\theta\}$ sur la valeur réellement observée x_0, on obtient l'intervalle de confiance Cx_0 pour θ.

Un avantage crucial des intervalles de confiance sur les tests de signification statistique est que les considérations liées à la taille des effets sont prises en compte. Dans l'exemple ci-dessus des PES, où la faible valeur-p se distinguait de la forte probabilité *a posteriori* de l'hypothèse nulle, les intervalles de confiance de 95% ou 99% pour θ auraient été de très étroits intervalles autour de θ_0. Autrement dit, dans le cas d'un échantillon de grande taille et d'un effet de petite taille, l'intervalle de confiance évite la fausse impression que l'hypothèse nulle est en tort et doit être rejetée.

Cependant, les intervalles de confiance ne peuvent être conseillés comme solution idéale. En premier lieu, l'idée que les scientifiques mettent au point des vrais *tests* en vue de vérifier l'adéquation de leur modèle statistique a complètement disparu, mais c'est un aspect vital et omniprésent de la pratique scientifique. En second lieu, les intervalles de confiance jouent davantage le rôle de tests de cohérence qu'ils n'inspirent confiance en une estimation particulière. Ils dressent la liste des ensembles de valeurs du paramètre étudié pour lesquelles les données obtenues n'auraient pas été rejetées à un seuil de $1 - \alpha$. C'est une perspective qui, dans son essence même, est pré-expérimentale. Mais cela

---

44 Voir par exemple : Geoff Cumming et Sue Finch, « Inference by Eye: Confidence Intervals and How to Read Pictures of Data », *American Psychologist*, vol. 60 / 2, 2005, p. 170-180.

ne garantit pas, après l'expérimentation, que le paramètre en question se situe « probablement » dans l'intervalle de confiance. C'est précisément le problème des statistiques fréquentistes de ne pas disposer d'une solide mesure de présomption post-expérimentale, problème qui pourrait, sur le long terme, faire pencher la balance en faveur de l'approche bayésienne.

4. Principes de vraisemblance et de règle d'arrêt

Les arguments précédents contre l'inférence fréquentiste présupposent tous, implicitement ou explicitement, une perspective bayésienne. Mais y a-t-il une manière de fonder l'inférence bayésienne sans supposer ce qui est à prouver ? Birnbaum est l'auteur d'une célèbre tentative en ce sens. Son argument est fondé sur le **principe de conditionnalité**, qui énonce que le soutien gagné par un mélange probabiliste d'expériences est égal à celui fourni par celle effectivement réalisée. Pour le dire de manière moins abstraite, supposons que l'on ait à choisir entre deux essais cliniques $\varepsilon_1$etε_2. Pour décider auquel procéder, on tire à pile ou face. La pièce tombe sur face et on procède uniquement l'essai ε_2. Birnbaum requiert alors que seules comptent comme preuves les informations obtenues en réalisant ε_2. Autrement dit, le fait qu'une autre expérience aurait pu être réalisée, si la pièce était tombée sur pile, ne devrait pas influencer notre raisonnement et nos conclusions.

Ce principe repose sur l'idée plausible selon laquelle tirer à pile ou face n'affecte aucunement le problème d'inférence que l'on cherche à résoudre. On peut donc conditionner sur le résultat du tirage à pile ou face. Si l'on admet que le principe de conditionnalité est plausible, on peut le combiner au principe de suffisance – un principe inoffensif d'inférence statistique que bayésiens et fréquentistes acceptent – pour dériver le

> **Principe de vraisemblance (PV).** Soit un modèle M comprenant un ensemble de mesures de probabilité $p(\bullet|\theta)$ paramétrées par $\theta \in \Theta$. Supposons que l'on réalise l'expérience ε dans M. Alors, la totalité du soutien à θ engendré par les

> preuves apportées par ε réside dans la fonction de vraisemblance $p(x|\theta)$, dans laquelle les données observées x sont considérées comme constantes[45].

Pour clarifier cette définition, précisons que la **fonction de vraisemblance** prend pour arguments les paramètres d'un modèle statistique et pour valeur la probabilité conditionnelle des données effectivement observées, sous l'hypothèse des valeurs de ces paramètres. En particulier le PV implique que la probabilité des résultats non observés n'affecte pas l'interprétation statistique d'une expérience (pour une discussion de ce principe et des différentes manières de le prouver, voir les ouvrages et articles de James Berger et Robert Wolpert[46], Deborah Mayo[47] et Greg Gandenberger[48]). Notons qu'un bayésien subjectiviste accepte automatiquement le PV : tout ce qui est requis pour passer d'une probabilité *a priori* à une probabilité *a posteriori* est la vraisemblance de H et de $\neg H$ en fonction des données observées. Dans un problème d'inférence statistique, cela correspond à la probabilité de x pour différentes valeurs du paramètre inconnu θ. Les fréquentistes, eux, considèrent comme pertinentes des informations en sus de la fonction de vraisemblance (comme par exemple la probabilité d'observer un résultat encore moins probable si l'on suppose l'hypothèse nulle) et sont donc en désaccord avec le principe de conditionnalité et le PV.

Le PV est plus qu'un principe purement théorique : il est vital pour interpréter les essais séquentiels en médecine. Dans ces situations, les

---

45 Allan Birnbaum, « On the foundations of statistical inference », *Journal of the American Statistical Association*, vol. 57 / 298, 1962, p. 269–306. James O. Berger et Robert L. Wolpert, *The Likelihood Principle*, Hayward, Institute of Mathematical Statistics, 1984.

46 James O. Berger et Robert L. Wolpert, *op. cit.*

47 Deborah G. Mayo, « An error in the argument from conditionality and sufficiency to the likelihood principle », in Aris Spanos, Deborah G. Mayo. *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science*, Cambridge, Cambridge University Press, 2010, p. 305-314.

48 Greg Gandenberger, « A New Proof of the Likelihood Principle », *The British Journal for the Philosophy of Science*, 2014.

règles d'arrêt définissent les conditions dans lesquelles doivent se terminer les tests portant sur l'efficacité d'un médicament. On peut, par exemple, mettre fin aux essais lorsque l'échantillon a atteint une certaine taille, ou lorsque les résultats soutiennent clairement une hypothèse plutôt qu'une autre. Les différends entre les défenseurs du PV et ses opposants, et entre bayésiens et fréquentistes, portent sur la question de savoir si une inférence, concernant par exemple l'efficacité d'un médicament, devrait être sensible à la règle d'arrêt utilisée.

Pour les adhérents au PV, les règles d'arrêt ne jouent aucun rôle dans la preuve de cette inférence. Berger et Berry[49] nomment cette exigence le **principe de la règle d'arrêt.** Leur justification est que :

> La conception d'une séquence d'essais expérimentaux est [...] ce que l'expérimentateur avait effectivement l'intention de faire.[50]

En d'autres termes, puisque de telles intentions sont « verrouillées à l'intérieur de la tête [de l'expérimentateur] »[51], invérifiables par d'autres expérimentateurs, et apparemment non connectées causalement aux processus qui génèrent les données, elles ne devraient pas affecter les inférences statistiques valides.

Cependant, d'un point de vue fréquentiste, certaines règles d'arrêt, comme celle qui prescrit un échantillonnage jusqu'à ce que les résultats favorisent clairement une hypothèse spécifique, nous mènent à des conclusions biaisées. En d'autres termes, la négligence d'une règle d'arrêt dans l'évaluation d'une expérience peut nous mener à une conclusion courue d'avance. Considérons une règle d'arrêt qui rejette l'hypothèse nulle ponctuelle $H_0 : \theta = \theta_0$ en faveur de $H : \theta \neq \theta_0$, lorsque les données

---

49 James O. Berger et Donald Berry, « The Relevance of Stopping Rules in Statistical Inference (with discussion) », in Shanti S. Gupta, James O. Berger. *Statistical Decision Theory and Related Topics IV*, New York, Springer, 1988, p. 29-72, p. 34.

50 Leonard J. Savage, *The Foundations of Statistical Inference*, Methuen, 1962, 122 p., p. 76. Voir aussi Ward Edwards, Harold Lindman et Leonard J. Savage, « Bayesian statistical inference for psychological research. », *Psychological Review*, vol. 70 / 3, 1963, p. 193, p. 239.

51 Leonard J. Savage, *op. cit.*, p. 76.

sont significatives pour un seuil de 5%. Cet événement arrivera à un certain moment avec une probabilité égale à 1, quelle que soit la valeur de θ[52]. De plus, il a été affirmé que certaines règles d'arrêtspratiquées en médecine (par exemple arrêter les essais cliniques lorsqu'un médicament s'est révélé plus efficace qu'un placebo) mènent à une surestimation peu plausible de la taille d'un effet expérimental.

D'un autre côté, il a été montré qu'aboutir à une conclusion courue d'avance n'est possible que si l'on utilise des mesures du soutien d'une hypothèse fréquentistes plutôt que bayésiennes. Autrement dit, on peut répondre à l'objection fréquentiste selon laquelle l'arrêt conditionnel introduirait un biais, que *ce n'est le cas que si l'on adopte une conception fréquentiste des preuves statistiques.* Il n'y a aucun problème si l'on combine l'interprétation bayésienne de l'arrêt conditionnel avec les évaluations bayésiennes des preuves[53]. Comme le débat concernant les mesures de soutien bayésiennes et fréquentistes, c'est-à-dire entre l'utilisation des valeurs-p et les facteurs de Bayes, la question méthodologique sur le rôle des règles d'arrêt dans les preuves semble être dans une impasse. Chaque camp de la controverse présuppose (ou doit présupposer) sa propre conception de l'inférence pour pouvoir critiquer la position adverse.

5. Discussion : la recherche de l'objectivité

Étant donnés tous les avantages du bayésianisme que nous avons énumérés jusqu'ici, pourquoi y a-t-il tant de résistance à l'introduction généralisée de méthodes bayésiennes ? Pourquoi les méthodes fréquentistes, avec tous leurs problèmes et bizarreries, dominent-elles la pratique scientifique ? On peut, d'après moi, identifier trois raisons. La première est qu'il y a des réserves de principe à l'encontre de l'approche

52 Leonard J. Savage, *op. cit.*. Deborah G. Mayo et Michael Kruse, « Principles of inference and their consequences », in *Foundations of Bayesianism*, Springer, 2001, p. 381–403.

53 Jan Sprenger, « Evidence and experimental design in sequential trials », *Philosophy of Science*, vol. 76 / 5, 2009, p. 637–649.

bayésienne car elle semble menacer l'objectivité, l'impartialité et l'autorité épistémique de la science. Bien que l'idéal d'une inférence statistique objective, libre de toute perspective personnelle, ait été sévèrement critiqué (par exemple par Heather Douglas[54] et Lorraine Daston & Peter Galison[55]) et ait pu perdre de son attrait pour de nombreux philosophes, il est toujours influent parmi les scientifiques et les instances de régulation, qui ont peur que des intérêts externes affectent les inférences. Pendant longtemps, des organismes comme la FDA[56] craignaient que l'analyse bayésienne puisse être utilisée à mauvais escient, pour éliminer de solides preuves scientifiques sur la base d'opinions *a priori*. La FDA ne s'est ouverte que récemment à l'analyse bayésienne des essais cliniques.

La deuxième raison du rejet de l'analyse bayésienne est que des institutions scientifiques comme les comités éditoriaux, les organismes de régulation et les associations professionnelles, sont inertes : elles ont tendance à s'en tenir aux pratiques qui ont « fait leurs preuves » et avec lesquelles elles sont familiarisées. Prenons l'exemple de la psychologie expérimentale : introduire ne serait-ce que quelques changements basiques, comme accompagner les valeurs-p d'estimation de la taille des effets et/ou du calcul de la puissance statistique, a été un processus long et difficile. Modifier les manuels et l'éducation des jeunes scientifiques pourrait prendre encore plus de temps. Un aspect plus positif est qu'un climat plus pluraliste qu'auparavant s'est installé ces dernières années, permettant à l'analyse bayésienne et à d'autres méthodes statistiques non-orthodoxes d'être l'objet d'un intérêt croissant.

La troisième raison du rejet du bayésianisme est que même certains spécialistes réputés des modèles bayésiens, comme Andrew Gelman et Cosma Shalizi[57], avouent que s'ils emploient les statistiques bayésiennes comme un outil technique, ils ne se décriraient pas eux-mêmes comme

---

54 Heather Douglas, *Science, Policy, and the Value-Free Ideal*, Pittsburgh, University of Pittsburgh Press, 2009.

55 Lorraine J. Daston et Peter L. Galison, *Objectivity*, Zone Books, 2007.

56 *Food and Drug Administration*, organisme états-unien chargé notamment de l'autorisation de mise sur le marché des médicaments, N.D.T.

subjectivistes. Leur approche méthodologique est plus proche de l'approche hypothético-déductive consistant à tester les modèles au moyen de leurs prédictions, de manière similaire à la justification fréquentiste de l'utilisation des tests d'hypothèses. Il semble donc que bien que les bayésiens semblent gagner haut la main d'un point de vue purement fondamental, il n'est pas évident que leurs méthodes constituent les meilleures solutions pour la pratique scientifique. Cela nous amène à narrer la façon dont l'inférence bayésienne est reliée au test de modèles statistiques dans un esprit hypothético-déductif, et plus généralement à étudier la relation entre des théories de la confirmation qualitatives et quantitatives, subjectives et objectives[58].

Le schéma d'inférence bayésien peut-il être modifié de manière à être plus objectif ? A cette question difficile, trois réponses, correspondant chacune à un programme de recherche différent, peuvent être esquissées.

Les probabilités objectives *a priori*. Comme l'explique Christian Robert dans le chapitre de ce volume sur les statistiques bayésiennes[59], il y a différentes manières de choisir une distribution de probabilité *a priori*. La méthode des probabilités objectives *a priori* cherche à enlever l'épine de l'accusation de subjectivisme du pied de l'analyse bayésienne, en acceptant des probabilités *a priori* qui appliquent une sorte de principe d'indifférence entre les hypothèses considérées (en attribuant par exemple à chaque valeur du paramètre une égale probabilité). Le problème de cette approche est que le principe d'indifférence qui lui est sous-jacent est philosophiquement fragile[60]. D'autres types de probabilités

57 Andrew Gelman et Cosma Rohilla Shalizi, « Philosophy and the practice of Bayesian statistics », *British Journal of Mathematical and Statistical Psychology*, vol. 66 / 1, 2013, p. 8–38.

58 Jan Sprenger, « A synthesis of Hempelian and hypothetico-deductive confirmation », *Erkenntnis*, vol. 78 / 4, 2013, p. 727–738.

59 Christian Robert, « Des spécificités de l'approche bayésienne et de ses justifications en statistique inférentielle », dans ce volume.

60 Voir à ce sujet Alan Hajek, « Interpretations of Probability », in Edward N. Zalta, (éd.). *The Stanford Encyclopedia of Philosophy*, éd. Edward N. Zalta, 2012.

a priori, comme celles motivées par l'invariance par transformation[61], sont plus prometteuses[62].

Le principe d'entropie maximale. Cette approche diffère de l'inférence bayésienne avec probabilités objectives *a priori,* en ce que le principe du conditionnement bayésien est rejeté. Les degrés de croyance rationnelle d'un agent devraient plutôt satisfaire trois contraintes[63] : ils devraient être soumis aux axiomes du calcul des probabilités, être conformes aux contraintes empiriques pesant sur nos degrés de croyances rationnels, et enfin ils devraient être **équivoques**, c'est-à-dire aussi proches de la moyenne que possible. Cette dernière contrainte revient à maximiser l'entropie de la distribution de probabilité en question. Si l'on note ω les « atomes » de la σ-algèbre concernée, l'entropie est donnée par le terme

$$H = -\sum_{\omega \in \Omega} p(\omega) \log p(\omega) \quad (7)$$

Si le principe d'entropie maximale est de grande aide dans de nombreux problèmes concrets d'ingénierie, d'informatique et d'autres disciplines proches, il est difficile de trouver une justification sans faille de l'idée que nos degrés de croyance devraient, en général, être aussi moyens que possible.

Conditionner sur le poids de la preuve. Des trois approches ici discutées, celle-ci est la moins connue. L'idée est de fournir une interprétation bayésienne valide des probabilités d'erreur fréquentistes en conditionnant de manière appropriée sur le poids de la preuve empirique[64].

---

61 Harold Jeffreys, *op. cit.*. José M. Bernardo, *op. cit.*

62 A ce sujet, pour une discussion des problèmes philosophiques sous-jacents, se reporter à Jan Sprenger, « The Renegade Subjectivist: José Bernardo's Reference Bayesianism », *Rationality, Markets and Morality*, vol. 3, 2012, p. 1–13.

63 Edwin T. Jaynes, « Prior probabilities », *Systems Science and Cybernetics, IEEE Transactions on*, vol. 4 / 3, 1968, p. 227–241. Jon Williamson, *op. cit.*

64 James O. Berger, « Could Fisher, Jeffreys and Neyman have agreed on testing? », *Statistical Science*, vol. 18 / 1, 2003, p. 1–32.

Le principe de conditionnement pourrait ainsi servir de pont entre bayésiens et fréquentistes. De plus, il est directement applicable à des problèmes saillants en analyse d'essais médicaux séquentiels[65]. De telles tentatives pour trouver un compromis entre inférences bayésiennes et fréquentistes sont toujours, pour la plupart, *terra incognita* d'un point de vue philosophique. Mais de mon point de vue, il y a beaucoup à gagner à étudier ces approches cherchant un terrain d'entente entre différentes écoles d'inférence statistique souvent décrites comme incompatibles.

Conclusion

Dans ce chapitre, on a introduit et comparé les inférences bayésienne et fréquentiste sous de nombreux aspects, en premier lieu à la lumière des mesures des preuves statistiques prônées par chacune. Nous avons vu que l'inférence fréquentiste souffre d'un certain nombre de problèmes conceptuels, épistémologiques et pratiques, qui semblent favoriser une approche bayésienne subjective. Cependant, lorsqu'il est question de l'objectivité des inférences statistiques et de rendre compte d'éléments cruciaux de la pratique scientifique (comme, par exemple, les tests d'hypothèse), le bayésien subjectiviste a du mal à joindre les deux bouts. Les recherches sur diverses formes d'inférence bayésienne objective poursuivies dans la littérature statistique représentent un programme de recherche fécond et enthousiasmant qui pourrait mettre à jour un terrain d'entente entre les deux grandes écoles d'inférence statistique qui soit à la fois philosophiquement solide et viable en pratique.

---

65 James O. Berger, Lawrence D. Brown et Robert L. Wolpert, « A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing », *The Annals of Statistics*, 1994, p. 1787–1807. Cecilia Nardini et Jan Sprenger, « Bias and conditioning in sequential medical trials », 2012.