Statistique et société
Novembre 2020

Volume 8, Numéro 2

Enseignement

Sommaire
Statistique et société

Volume 8, Numéro 2

Sommaire
Statistique et société

Volume 8, Numéro 2

Varia

Recensions

Statistique et société

Magazine quadrimestriel publié par la Société Française de Statistique.
Le but de Statistique et société est de présenter, d'une manière attrayante et qui invite à la réflexion, l'utilisation pratique de la statistique dans tous les domaines de la vie. Il s'agit de montrer comment l'usage de la statistique intervient dans la société pour y jouer un rôle souvent inaperçu de transformation, et est en retour influencé par elle. Un autre dessein de Statistique et société est d'informer ses lecteurs avec un souci pédagogique à propos d'applications innovantes, de développements théoriques importants, de problèmes actuels affectant les statisticiens, et d'évolutions dans les rôles joués par les statisticiens et l'usage de statistiques dans la vie de la société.

Design graphique
fastboil.net

ISSN 2269-0271

Éditorial

Emmanuel DIDIER
Rédacteur en chef de Statistique et société

Chère lectrice, cher lecteur,

Ce numéro est le premier dont le dossier est consacré à l'enseignement de la statistique, un sujet qui nous tient à cœur. En effet, *Statistique et société* a fusionné il y a quelque temps avec *Statistique et enseignement*, une autre revue de la SFdS. Nous en étions très heureux car il est évident qu'un domaine où la statistique et la société sont en interaction forte et riche est bien celui de la formation des futurs citoyens et des praticiens.

Ce numéro montre que les potentiels d'enrichissement mutuels entre les deux revues se sont réalisés. Je voudrais donc aujourd'hui rendre un hommage tout particulier à Catherine Vermandele et Antoine Rolland qui ont rejoint le comité de rédaction, s'y sont merveilleusement intégrés et à qui nous devons ce dossier. D'autres publications portant sur ce thème ne manqueront pas de suivre.

L'enseignement est abordé d'abord sous forme d'un compte rendu d'expérimentation apte à convaincre les élèves de l'intérêt des probabilités, écrit par Eric Parent, Jean-Jacques Boreux, Etienne Rivot et Sophie Ancelet. Suit un article d'Annika Näslund sur un nouveau master européen formant les futurs statisticiens publics. Vient ensuite le compte-rendu par Alexis Eidelman d'un manuel d'utilisation du langage Python. On le voit, l'enseignement ici n'est pas envisagé dans une classe grise et froide, avec craie qui crisse sur le tableau noir et tablier bleu terne. Non, l'enseignement ici est joyeux, porte sur des innovations et vise à construire le futur.

Un article *varia* de Luc Arrondel, Richard Duhautois et Jean-François Laslier répond enfin à une question que tout amateur de foot se pose : est-ce un avantage de commencer les tirs au but ? Je ne dévoile pas la réponse apportée par un texte qui s'avale aussi facilement qu'une bonne bière à la mi-temps.

Viennent ensuite trois recensions d'ouvrages qui montrent combien l'actualité éditoriale est riche aussi en ce qui concerne les rapports entre la statistique et la société. Antoine Rolland nous présente une réédition d'un ouvrage de W.E.B. Du Bois sur le décompte des noirs aux USA au début du XX^{ème} siècle. Thomas Amossé présente la dernière livraison de Walter Radermacher, infatigable et créatif ancien directeur d'Eurostat, portant sur le rôle des statistiques publiques dans un monde transformé par les data. Enfin, Gérard Lang rend compte des réflexions de Jean-Pierre Le Gléau sur l'importante question du secret statistique.

Bonne lecture !
Emmanuel Didier

Une expérience ludique de capture-marquage-recapture pour l'initiation au raisonnement probabiliste indispensable au statisticien-modélisateur

Éric PARENT[1]
AgroParisTech, UMR 518, Paris

Jean-Jacques BOREUX[2]
Dpt Sc. & G. Environnement, Université de Liège, site d'Arlon

Étienne RIVOT[3]
UMR ESE, Ecology and Ecosystem Health, Institut Agro, INRAE, Rennes

Sophie ANCELET[4]
Institut de Radioprotection et de Sûreté Nucléaire (IRSN), PSE-SANTÉ/SESANE/LEPID, Fontenay-Aux-Roses

TITLE

A ludic experience of capture-mark-recapture for the initiation to probabilistic reasoning essential to the statistician-modeller.

RÉSUMÉ

Les méthodes de capture-marquage-recapture sont des méthodes astucieuses d'échantillonnage peu invasives pour évaluer le nombre d'individus dans une population. Utilisées principalement en écologie, elles trouvent aussi des applications de portée bien plus large dans divers domaines tels que la sociologie et la psychologie expérimentales. Du point de vue de la pédagogie, elles permettent d'illustrer de façon simple, pratique et vivante de nombreux points-clés du raisonnement probabiliste indispensables au statisticien-modélisateur. À l'aide d'une expérience ludique facile à effectuer en salle avec des gommettes, des haricots secs, une cuillère à soupe et un saladier, nous montrons comment aborder de façon simple et intéressante les points-clés suivants dans le cadre d'un problème d'estimation de la taille inconnue d'une population :

– les ingrédients de base du problème de statistique inférentielle considéré, en particulier, inconnues versus observables ;

– la construction d'un modèle probabiliste/stochastique possible, fondé sur l'assemblage de plusieurs briques binomiales élémentaires, ainsi que les différentes décompositions possibles de la vraisemblance associée ;

– la recherche d'estimateurs, leur étude théorique ainsi que la comparaison de leurs propriétés mathématiques par simulation numérique ;

1. eric.parent@agroparistech.fr
2. jj.boreux@ulg.ac.be
3. etienne.rivot@agrocampus-ouest.fr
4. sophie.ancelet@irsn.fr

– les différences opérationnelles majeures entre approches statistiques fréquentielle et bayésienne. Cette expérience permet également d'illustrer en quoi le travail d'un statisticien-modélisateur ressemble bien souvent à celui d'un enquêteur de police…

Mots-clés **:** *capture-marquage-recapture, estimation, loi binomiale, raisonnement probabiliste, statistique bayésienne.*

ABSTRACT

Capture-mark-recapture techniques are smart non-invasive sampling methods to evaluate the number of individuals in a population. Primarily used in ecology, they also find applications with a much broader scope in various fields such as experimental sociology and psychology. From a pedagogical standpoint, they nicely illustrate, in a very simple and practical way, many key points of probabilistic reasoning as a rationale essential to statistician-modellers. Starting from an affordable toy experiment that can be easily performed indoors with stickers, beans, a tablespoon and a saladbowl, we show how to deal with the following key points in a simple and interesting way, in the specific context of estimating the unknown size of a population:
– the basic ingredients of a problem of inferential statistics, unknowns versus observables;
– the design of a possible probabilistic/stochastic model by assembling several binomial building blocks, as well as the many possible decompositions of the associated likelihood;
– the search for estimators, their theoretical study as well as the comparative study of their mathematical properties using computer simulations;
– the major operational differences between the frequentist and Bayesian statistical approaches. This experiment also illustrates how the everyday work of a statistician-modeller often resembles the one of a police investigator…

Keywords**:** *capture-mark-recapture, estimation, binomial distribution, probabilistic reasoning, Bayesian statistics.*

1. Introduction

Les stats ? Un pensum ! L'enseignement de la statistique peut avoir laissé un souvenir tristement aride, et ce, même pour les anciens étudiants les plus passionnés par les mathématiques, d'ailleurs peut-être à leur tour devenus enseignants-chercheurs en statistique et/ou statisticiens dans le secteur public ou privé. Depuis le siècle dernier, l'approche pédagogique la plus souvent privilégiée en statistique inférentielle consiste à consciencieusement faire reconnaître par les étudiants *une* situation dans un catalogue de différentes situations typiques afin d'appliquer, par analogie, *la* technique statistique appropriée tirée d'une boîte à outils mathématiques. Ainsi, par exemple, l'estimation de l'effet de covariables sur une variable-réponse binaire appelle souvent systématiquement les étudiants à l'utilisation d'une régression logistique et celle d'une variable-réponse de type comptage à l'utilisation d'une régression de Poisson. À noter que ces outils sont, pour la plupart, déjà implémentés de manière générique et optimale dans les logiciels de statistique classiques (R, SAS, Stata, ...), ce qui facilite largement leur utilisation mais peut parfois nuire à la réflexion critique.

Aujourd'hui, le fraîchement émoulu *data scientist* reçoit par ailleurs un enseignement axé sur les défis informatiques qui se posent lorsqu'on essaie d'utiliser des données massives pour répondre à un questionnement. Pour autant, la compréhension des points-clés du raisonnement probabiliste – permettant la mise en équations explicite d'un problème concret en vue de faire parler des données – aura-t-elle réellement progressé ces dernières années ? Bien que les modèles stochastiques les plus complexes (décrits dans un cours de Master par exemple) ne reposent, le plus souvent, que sur l'assemblage de sous-modèles élémentaires beaucoup plus simples (Wikle et al., 1998), force est de constater que de nombreux maîtres de stage déplorent souvent par la suite la timidité excessive, voire le manque d'autonomie et d'envie créatrices des apprentis-chercheurs qu'ils accueillent, notamment en début de thèse. À ce niveau académique, les compétences de formalisation sont pourtant indispensables : faute de réaliser l'éventail des modélisations possibles que permettent déjà les structures aléatoires de base, comment contribuer au développement de nouveaux modèles ?

La responsabilité de ce handicap amène à se questionner sur la pédagogie que nous-mêmes déployons lors de nos enseignements de niveau M2 ou inférieur. En ces temps où nombre de tenants de l'*Intelligence Artificielle* et du *Big Data* plaident volontiers pour un apprentissage sans modèle statistique, comment motiver notre audience à saisir l'intérêt de la construction explicite d'un raisonnement probabiliste pour répondre à un problème concret ?

La mise en place d'expériences ludiques et faciles à réaliser en salle par les étudiants peut permettre de rendre tangibles et intéressantes la construction explicite et les propriétés mathématiques de modèles probabilistes, qu'ils soient élémentaires ou déjà élaborés (*e.g.*, hiérarchiques). Ainsi, afin d'illustrer la construction du modèle de Bernoulli pour tests en duo/trio, Azaïs (2004) décrit une expérience ludique de dégustation pour distinguer deux produits semblables (sodas au cola) à partir de résultats binaires. De même, afin d'illustrer la construction de la loi normale multivariée, Tibshirani et al. (2011) modélisent les impacts au jeu de fléchettes, rappellent les propriétés de la transformée de Fourier et cherchent la stratégie judicieuse pour viser à maximiser son score. La dernière section du livre *Teaching Statistics* de Gelman et Nolan (2017) est consacrée à un atelier de construction d'hélicoptères en papier (Box, 1992; Annis, 2005), afin de mettre en place, ciseaux en main, une séance très appliquée de planification expérimentale.

Les manuels de statistique ont tendance à laisser croire que la loi de Bernoulli n'est à réserver qu'aux premières séances d'un cours élémentaire de statistique. En accord avec Collett (2002), nous réfutons ce parti pris. Cette loi de probabilité est d'intérêt dans de nombreux cursus comme les sciences de la vie ou les sciences humaines et sociales. Elle trouvera également son intérêt dans le cadre d'une initiation au raisonnement probabiliste, qui peut se faire, par exemple, dans un cours de statistique portant sur l'approche bayésienne et/ou la modélisation hiérarchique.

Dans cet article, nous développons les idées en germe dans Dudley (1983) qui illustraient une ex-

périence de capture-marquage-recapture (CMR) avec des friandises colorées du type *M&M's*. Dans la section 2, nous décrivons les détails d'une expérience ludique facile à effectuer en salle avec des gommettes, des haricots secs, une cuillère à soupe et un saladier. Nous l'avons réalisée de nombreuses fois avec des étudiants de niveau M1 d'un cursus de mathématique, de niveau M2 d'un cursus d'écologie, et de dernière année d'écoles d'ingénieurs en statistique, dans le cadre d'ateliers d'enseignement et d'écoles-chercheurs. Si on les livre à eux-mêmes lors d'une première demi-heure, les étudiants, matheux ou écologues, se précipitent généralement sur l'obtention d'une valeur ponctuelle, sans se soucier d'indiquer la moindre variabilité. Tous sont enclins à confondre probabilité et fréquence empirique. Certains écologues sont peut-être un peu moins réticents à admettre la présence d'incertitudes, mais aucun étudiant n'a d'emblée recours à un vocabulaire probabiliste. En réaction, nous pensons que prendre le temps d'exploiter tous les aspects de cette expérience permet de mettre en lumière de façon simple et intéressante les points-clés du raisonnement probabiliste, indispensables au statisticien-modélisateur, dans le cadre spécifique d'un problème d'estimation de la taille inconnue d'une population :

– les ingrédients de base du problème de statistique inférentielle considéré, en particulier en distinguant les grandeurs que l'on voit, les observables, de celles qu'on ne voit pas mais qui sont nécessaires pour poser le problème, les inconnues (section 3) ;

– la construction d'un modèle probabiliste possible, basé sur l'assemblage de plusieurs briques binomiales élémentaires, ainsi que les différentes décompositions possibles de la vraisemblance associée (section 4) ;

– la recherche d'estimateurs, leur étude théorique ainsi que la comparaison de leurs propriétés mathématiques par simulation numérique, qui feront l'objet de la section 5 ;

– les différences opérationnelles majeures entre approche statistique fréquentielle et perspective bayésienne, discutées dans la section 6.

2. Une expérience ludique de capture-marquage-recapture à réaliser en groupes

Le recensement est le dénombrement exhaustif des individus constituant une population statistique. Dans la pratique, un recensement nécessite généralement des moyens importants et la durée des enquêtes est un frein quand l'estimation de l'effectif doit être rendue rapidement. Dans ce contexte, travailler sur un modèle probabiliste de capture-marquage-recapture pour estimer la taille d'une population est un recours intéressant. Pour pouvoir s'inscrire dans les hypothèses des modèles standards de la famille CMR, il faut notamment, comme pour un recensement, que le nombre d'individus dans un milieu fermé n'évolue pas : les mouvements migratoires, les processus naturels de natalité et de mortalité sont tous nuls ou s'équilibrent durant la mesure. Il faut par ailleurs que la répartition des individus dans le milieu soit homogène et que la probabilité de capture ne varie pas durant la durée des opérations.

Historiquement, ce sont les écologues intéressés par l'estimation de la taille des populations animales dans le milieu naturel qui ont développé le modèle CMR (Seber et al., 1982; McCrea et Morgan, 2014; Royle et al., 2013). Nous prenons ici l'exemple de l'estimation d'une population de poissons vivant dans un milieu fermé, comme un lac ou un tronçon de rivière délimité en amont et en aval par des filets. Le client du statisticien pourrait être la gestionnaire d'une société de pêche intéressée par le nombre de saumons juvéniles quittant une rivière pour rejoindre la mer comme dans Rivot et Prévost (2002). Une première pêche (considérée comme une première opération de capture) donne C_1 poissons qui sont marqués et remis à l'eau. Une fois l'homogénéité du milieu

restaurée, une seconde pêche (souvent avec le même dispositif) fournit C_2 poissons dont C_{21} poissons qui ont été marqués lors de la première pêche et donc $C_{20} = C_2 - C_{21}$ poissons qui sont non marqués. Le modèle CMR trouve également application dans d'autres domaines, en particulier au cours d'expérimentations pour les sciences humaines et sociales. Par exemple, Leyland et al. (1993) relatent comment la ville de Glasgow a mis en œuvre une méthode CMR pour évaluer le nombre de prostituées afin de mieux appréhender les risques de propagation des maladies infectieuses.

Sauf rares exceptions, il n'est pas possible d'emmener toute une classe sur le terrain pour collecter des données de ce type. Aussi, nous proposons une expérience de CMR simple à réaliser en salle et en groupes, en vue d'estimer la taille inconnue d'une population. Afin de rendre la mise en situation la plus concrète possible, nous utiliserons donc par la suite le vocabulaire emprunté à l'écologie halieutique : *pêche* pour *capture*, *poisson* pour *individu*, etc.

FIGURE 1 – *Gommettes, haricots secs, cuillère à soupe et saladier lors d'une expérience de capture-marquage-recapture réalisée en salle*

Cette expérience de CMR nécessite que les étudiants soient répartis en groupes ; nous avons constaté qu'un effectif de trois étudiants par groupe permet un partage du travail efficace et une discussion inventive. L'expérience proposée a l'avantage de ne demander que peu de matériel à mettre à disposition de chaque groupe (voir la figure 1) :

- une cuillère à soupe : c'est l'instrument de pêche ;
- un saladier (diamètre 30 à 40 cm) ou un moule à cake ($30 \times 10 \times 10$ cm^3) : c'est le lac (ou le tronçon de rivière) ;
- un kilogramme de riz pour simuler l'eau du lac ;
- un paquet de haricots secs, 500 grammes de lingots blancs ou rouges feront l'affaire : ce sont les poissons ;
- Un paquet d'étiquettes auto-adhésives (*e.g.*, gommettes) de couleurs variées pour le marquage des poissons.

L'expérience se déroule comme suit. Tout d'abord, chaque groupe d'étudiants prépare le saladier du groupe voisin : il y dispose une couche de riz ainsi qu'un certain nombre de haricots secs qu'il aura

préalablement comptés. Requérant la discrétion des préparateurs, le professeur relèvera un à un les effectifs exacts [5] de la population contenue dans chacun des saladiers (*i.e.* l'état de la nature). Après avoir récupéré son saladier auprès de son voisin (libre au professeur d'imaginer de plus savantes permutations), chaque groupe doit réaliser, au moins dans un premier temps, deux pêches successives, la taille de la population qu'il évalue lui étant donc inconnue. En pratique, il est conseillé de faire au moins six coups de cuillère par pêche afin d'obtenir des captures suffisamment nombreuses pour une bonne estimation des inconnues du problème (décrites dans la section suivante). Après chaque pêche, le groupe :

a) marque les poissons pêchés à l'aide de gommettes d'une couleur fixée mais différente d'une pêche à l'autre (voir la figure 2) ;

b) compte et note le nombre de poissons pêchés.

Nous suggérons également de marquer la seconde pêche et d'effectuer une troisième pêche, bien que ce ne soit guère l'usage rencontré sur le terrain. Cette expérience permet ainsi à chaque groupe de récolter ses propres données de CMR.

FIGURE 2 – *Étudiants en action lors d'une expérience de capture-marquage-recapture*

Pour chaque groupe d'étudiants, l'objectif est de proposer une estimation du nombre de haricots secs contenus dans son saladier **et** une fourchette de confiance, au vu des données qu'il a collectées. Outre observer le mode de fonctionnement en collectif des étudiants, il est intéressant de leur faire exprimer leur conception de ce qu'est une probabilité, et comment s'y prendre pour répondre à l'objectif fixé. Face à un problème réel, comment ont-ils recours au raisonnement probabiliste et aux outils statistiques ? Comment assimilent-ils les informations ? Quoique les questions posées soient très simples : *Combien de poissons ? Avec quelle (in)certitude ?*, le problème probabiliste est déjà suffisamment élaboré pour donner matière à un questionnement scientifique fructueux.

5. On supposera que ces effectifs auront été comptés séparément par chacun des 3 étudiants afin de limiter au maximum le risque d'erreur de mesure

3. Les ingrédients de base d'un problème de statistique inférentielle

Il n'est certes pas facile de réfréner l'impatience amusée des étudiants pour pêcher des haricots et coller des gommettes, mais on peut profiter de leur intérêt pour les faire réfléchir aux résultats qu'ils vont obtenir au terme de leur expérience de CMR. Pour plus de clarté, on se focalisera plutôt, à ce stade, sur le cas simple de deux pêches successives. La maïeutique prescrit de demander aux étudiants de lister puis de classer l'ensemble des ingrédients qui vont jouer un rôle dans le problème de statistique inférentielle posé. Pour faire progresser les idées, il faut oser utiliser des mots dont le sens se précisera au fur et à mesure de leur emploi dans divers contextes, car c'est en remettant régulièrement l'ouvrage sur le métier que l'on affine sa compréhension. La notion de probabilité en est un parfait exemple. Nous suggérons ici de commencer par rappeler aux étudiants qu'en statistique inférentielle, deux sortes d'objets se distinguent selon leur nature et qu'il faut les nommer pour avancer.

3.1. Les observables

Les observables désignent les grandeurs qu'on peut voir, toucher ou mesurer. Par convention, une lettre majuscule latine est utilisée pour les nommer lorsqu'il s'agit de variables aléatoires. A ce stade, un rappel informel simple de ce que désigne une variable aléatoire en théorie des probabilités peut être nécessaire, comme par exemple : *il s'agit d'une quantité inconnue, avant d'avoir réalisé l'expérience de CMR, et qui peut prendre une collection imaginée de valeurs, munies de pondérations*.

Dans le cas d'une méthode de CMR fondée sur deux pêches successives, les observables C_1, C_{20}, C_{21} et C_2 ont déjà été définies à la section 2. Un petit dessin valant mieux qu'un long discours, nous pensons pertinent de présenter en complément le diagramme de Venn de la figure 3 sur lequel sont notamment représentés ces différents résultats de comptage.

Une observable est bien sûr à distinguer de son observation, sa réalisation qui, elle, est associée à une valeur numérique unique – stockable dans un ordinateur (par exemple, après avoir réalisé une expérience de CMR). Par convention, une observation est nommée avec une lettre minuscule latine.

3.2. Les inconnues

Les inconnues désignent les grandeurs qu'on ne voit pas. A ce stade, il est important de modérer une discussion générale pendant laquelle il faut insister constamment sur le fait que nous faisons *exister* les inconnues parce qu'on nomme ces concepts qui sortent de notre imagination. Il faut également faire astucieusement miroiter une facilité d'emploi à venir pour justifier l'usage de choisir des lettres grecques[6] pour les désigner. Après un temps de discussion apparaissent infailliblement les paramètres des lois d'échantillonnage de notre problème de statistique inférentielle :

- ν : la taille de la population ;
- π_1 : la probabilité de capture au cours de la première pêche ;
- π_2 : la probabilité de capture au cours de la seconde pêche, éventuellement supposée égale à π_1.

Dans notre problème, c'est ν le véritable paramètre d'*intérêt* car c'est l'inconnue sur laquelle se focalisera principalement l'écologue. L'introduction de π_1 et π_2 dans le raisonnement probabiliste formel est néanmoins indispensable pour estimer ν.

6. A cette occasion, nous nous sommes aperçus que, face à un public de plus en plus mondialisé, nous ne pouvions souvent plus tenir comme allant de soi la connaissance partagée de la culture gréco-latine.

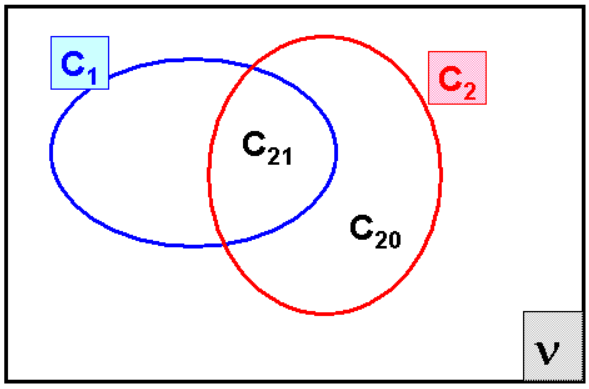


FIGURE 3 – *Diagramme de Venn indiquant les différents résultats de comptage obtenus lors d'une méthode de capture-marquage-recapture basée sur deux pêches successives. C_1 est le nombre total de poissons pêchés et marqués la première fois, C_{21} désigne le nombre de poissons pêchés lors de la seconde pêche ayant déjà été marqués lors de la première pêche, C_{20} le nombre de poissons pêchés lors de la seconde pêche non marqués lors de la première pêche.*

4. Construction d'un modèle probabiliste par assemblage de briques élémentaires binomiales

4.1. Passer des inconnues aux observables

L'expérience est toujours en attente mais inconnues et observables ayant été identifiées, on peut désormais enquérir les étudiants du « moyen » pour passer des premières aux secondes, comme sur la figure 4 sur laquelle on a pris soin de faire figurer les inconnues en haut (le monde éthéré des abstractions) et les observables en bas, afin de désigner le niveau du terrain de la réalité expérimentale. À un certain moment apparaît un consensus pour construire un modèle probabiliste. Il est alors profitable d'insister sur l'idée d'un assemblage de plusieurs lois binomiales élémentaires (appelées briques par la suite, par analogie avec les jeux de Lego) qui miment les résultats de capture et de recapture d'une expérience basée sur deux séquences successives de pêche. Le résultat C_1 de la première pêche correspond au nombre de poissons pêchés dans un lac ou un tronçon de rivière (délimité en amont en aval par des filets) contenant ν poissons. Sous les hypothèses décrites dans la section 2, une loi binomiale de paramètres ν et π_1 semble être un choix pertinent possible : chaque poisson est capturé de manière indépendante et avec la même probabilité π_1. Suivant la même logique, les résultats C_{21} et C_{20} de la deuxième pêche peuvent aussi être modélisés à l'aide de lois binomiales mais dont le paramètre de probabilité est π_2 et le nombre total de tirages possibles est C_1 et $\nu - C_1$ respectivement (Figure 3). Un point intéressant à souligner est l'assemblage des lois binomiales d'une pêche à l'autre : le résultat de la première pêche conditionne le paramétrage de la loi binomiale de la seconde pêche.

Est-ce que cette étape de construction explicite du modèle est un passage obligé ? Étonnamment, cette formalisation mathématique est également perçue comme un moyen de rechercher un agré-

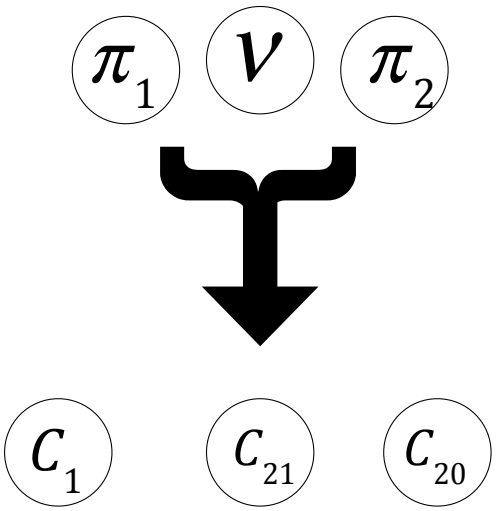


FIGURE 4 – *Faire un modèle ? Passer des inconnues aux observables !*

ment du collectif, voire une garantie de sérieux scientifique.

A ce stade, il nous semble utile de veiller à insister sur les trois points suivants, qui constituent des préalables indispensables pour répondre à tout problème concret de statistique inférentielle.

(i) Pouvoir lister et justifier ses hypothèses de modélisation – Il faut en faire pour avancer et limiter la gamme des incertitudes qu'on accepte de représenter. Mais lesquelles ? Par exemple, il faut rendre parlantes les conditions selon lesquelles les pêches sont indépendantes et suivent la même distribution de probabilité : *les poissons ne se parlent pas, ils se font capturer de la même façon quelle que soit leur taille, il n'y pas d'effet du marquage sur le comportement des poissons, etc.* Ces hypothèses sont-elles réalistes et objectives ? Pourquoi les fait-on ? Il est facile de partager l'avis selon lequel $\pi_1 = \pi_2$. Cette probabilité commune sera notée π par la suite. Par contre, une hypothèse qu'on justifie par sa commodité, comme, par exemple : « *Techniquement, c'est uniquement ce qu'on saura traiter et il nous faut bien modestement commencer par quelque chose.* » est une potion qui reste encore trop amère à avaler pour certains étudiants, nourris de déterminisme avec l'idée d'une science qui ne saurait être qu'exacte. N'ont-ils pas compris qu'un résultat scientifique est toujours né d'une simplification du problème posé, ce qui revient à considérer un cadre formel duquel on ne se donne pas la permission de sortir avant de finir le raisonnement et, alors seulement, éventuellement le remettre en question ?

(ii) Pouvoir simuler des données grâce à un programme informatique – En tant que professeur, nous voudrions poursuivre l'idée d'un assemblage de plusieurs briques binomiales élémentaires du type :

$$\begin{aligned} C_1|\nu,\pi_1 &\sim dbin(\pi_1,\nu),\\ C_{20}|C_1,\nu,\pi_2 &\sim dbin(\pi_2,\nu-C_1),\\ C_{21}|C_1,\pi_2 &\sim dbin(\pi_2,C_1), \end{aligned} \tag{1}$$

en utilisant, par exemple, la syntaxe *dbin* de type BUGS (Lunn et al., 2000) pour référer à la loi binomiale. Sauf rares exceptions, l'étudiant ne réagit pas immédiatement de cette façon qui lui semblera trop éthérée. En revanche, il ne renâclera pas à écrire et à exécuter sur son ordinateur un programme informatique de simulation de l'expérience. Reste à faire comprendre qu'appeler des fonctions de la famille *random* dans un algorithme, c'est se trouver *de facto* en présence d'un modèle probabiliste. Ceci est le message important à faire passer (et heureusement facile compte-tenu de l'engouement pour l'ordinateur). Pour notre cas d'étude, une routine de simulations de résultats de CMR pourrait par exemple s'écrire en R comme suit :

```
nu <- 1000
pi1 <- 0.75
pi2 = pi1
C1 = rbinom(1,nu,pi1)
C21 = rbinom(1,C1,pi2)
C20 = rbinom(1,nu-C1,pi2)
C2 = C21+C20
```

A noter que, pour simuler des réalisations des observables (*i.e.* produire des données), on doit se placer dans la situation où les inconnues sont connues (ou supposées connues), ce qui est le cas de ν , π_1 et π_2 dans l'exemple ci-dessus.

(iii) Construire un graphe acyclique orienté – Une bonne idée pour aider à poser les hypothèses de modélisation avant les conclusions est d'introduire un graphe acyclique orienté (voir par exemple Spiegelhalter et al. (1993)). Un tel graphe est souvent désigné par l'acronyme anglais *DAG*, pour *Directed Acyclic Graph*. Un DAG s'appuie sur des règles graphiques simples qui permettent d'utiliser des images intuitives pour aider à la conception de modèles et faciliter leur présentation . Les cercles (ou nœuds) représentent des variables aléatoires (*i.e.* inconnues ou observables), les flèches en traits pleins traduisent des lois de probabilité conditionnelles et les flèches en pointillés des opérations arithmétiques intermédiaires, comme sur la représentation de la figure 5 : $\nu - C_1$ désigne le nombre de poissons non marqués après la première pêche, mais aussi l'un des paramètres de la loi binomiale suivie par la variable aléatoire C_{20}.

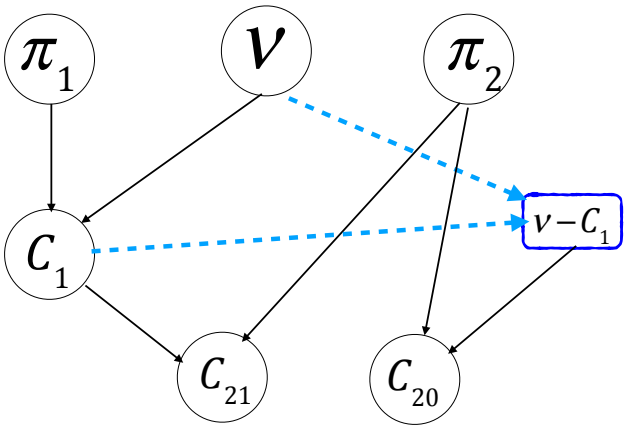


FIGURE 5 – *Une bonne idée pour aider à la conception d'un modèle probabiliste et faciliter sa présentation : un graphe acyclique orienté*

4.2. Tous les chemins mènent à Rome

Construire un modèle probabiliste paramétrique pour décrire une expérience de CMR revient à imaginer de multiples tirages aléatoires en cascade dans des urnes (voir la section 4.1), avec pour objectif de reproduire *in fine* des nombres comparables aux observations réellement faites. Comme on va le mettre en évidence ci-après, l'exemple CMR a pour avantage pédagogique de permettre de montrer qu'on peut parfois atteindre le même but par des chemins distincts, mais qui conduisent à la même vraisemblance et donc au même modèle probabiliste paramétrique.

Bien sûr, construire un DAG ne semble pas à proprement parler équivalent à écrire une vraisemblance, un mot-clé du cours de statistique que les étudiants ont généralement tous retenu. Hélas, ces derniers ne maîtrisent pas toujours les notions apportées par cette hydre à deux têtes : s'agit-il d'une fonction des données, ou bien des paramètres ? Une façon de remettre les notions en place

est d'adopter la notation « crochets » de Gelfand et Smith (1990). Nous suggérons de définir d'abord la loi d'échantillonnage de C_1, C_{20}, C_{21} comme $[C_1, C_{20}, C_{21}|\nu, \pi_1, \pi_2]$ pour souligner qu'il faut se mettre en situation d'une puissance créatrice qui connaît les inconnues ν, π_1 et π_2 pour écrire le modèle[7]. Mais quand on voit cette même expression mathématique comme une fonction des inconnues (son second argument), il s'agit alors de la vraisemblance du modèle. En effet, le terme rebattu *maximum de vraisemblance* n'indique-t-il pas que c'est alors une fonction des inconnues ?

À ce stade, les propositions des étudiants mettront en évidence plusieurs façons de décrire un même modèle CMR. Ces propositions sont à réorganiser après avoir aidé à écrire chacune d'entre elles de façon un peu plus formelle.

La première possibilité est de suivre un raisonnement probabiliste conditionnel séquentiel. En d'autres termes, il s'agit d'opter pour la vision constructive séquentielle rencontrée dans la section 4.1 qui, fondée sur des hypothèses d'indépendance conditionnelle, permet d'écrire la loi de probabilité jointe des observables C_1, C_{20} et C_{21} sachant ν, π_1 et π_2 comme la décomposition en produits suivante :

$$[C_1, C_{20}, C_{21}|\nu, \pi_1, \pi_2] = [C_1|\nu, \pi_1] \times [C_{20}|C_1, \nu, \pi_2] \times [C_{21}|C_1, \pi_2].$$

À noter que le DAG de la figure 5, qui laisse apparaître clairement les relations d'indépendance conditionnelle entre observables, peut considérablement aider dans l'écriture de cette décomposition. On suit alors quasi-mécaniquement les deux étapes de l'expérience de CMR. La décomposition ci-dessus ainsi que les briques binomiales décrites dans la section 4.1 justifie le modèle décrit par le jeu d'équations (1).

Comme dans le chapitre 5 de Marin et Robert (2007), il peut être intéressant d'opposer à cette vision séquentielle la vision globale qui établit directement le bilan d'une expérience de CMR à deux pêches successives, à l'aide d'une loi multinomiale à 4 catégories[8] et dont les probabilités et effectifs associés sont indiqués dans le tableau 1.

Tableau 1 – *Probabilités de capturer un poisson et effectif associé à chacune des 4 catégories possibles (listées en colonne) d'une loi multinomiale décrivant le bilan global d'une expérience de CMR fondée sur deux pêches successives*

	Pêches 1 et 2	1re pêche seulement	2e pêche seulement	Jamais
Probabilité	$\pi_1\pi_2$	$\pi_1(1-\pi_2)$	$(1-\pi_1)\pi_2$	$(1-\pi_1)(1-\pi_2)$
Effectif	C_{21}	$C_1 - C_{21}$	C_{20}	$\nu - (C_1 + C_2 - C_{21})$

Enfin, certains souhaiteront profiter de la séance pour faire un rappel de calcul de probabilités en demandant tout d'abord aux étudiants de vérifier que la loi de $C_1, C_{20}, C_{21}|\nu, \pi_1, \pi_2$ est la même sous le modèle binomial séquentiel et le modèle multinomial de bilan. Dans le texte du matériau supplémentaire, on montre cette première équivalence puis, comment on retombe, par un simple changement de variable, sur la modélisation traditionnelle alternative d'une expérience de CMR, qui suppose un tirage sans remise des poissons recapturés lors de la 2e pêche via une loi hypergéométrique[9] :

7. Reconnaissons-là un tropisme bayésien partagé par les auteurs : la sigma-algèbre qui permettrait de définir une distribution conjointe sur C_1, ν, π_1, C_{21}, C_{20} et π_2 est ici enfouie à cent lieux sous le tapis.
8. Il faut ici avoir soin de ne pas postuler trop rapidement l'élimination de l'indice du π afin d'aider à distinguer facilement les deux phases de pêche.
9. Rappelons l'interprétation classique d'une variable aléatoire Y de loi hypergéométrique $(N+B, N, K)$: après un tirage sans remise de K boules dans une urne contenant N boules noires et B boules blanches, on observe le nombre y de boules noires obtenues. Sa loi de probabilité s'écrit :

$$[Y = y|N+B, N, K] = \frac{C_N^y C_B^{K-y}}{C_{N+B}^K}.$$

$$\begin{aligned} C_1|\nu,\pi_1 &\sim dbin(\pi_1,\nu), \\ C_2|\nu,\pi_2 &\sim dbin(\pi_2,\nu), \\ C_{21}|C_1,C_2,\nu &\sim hypergeometrique(\nu,C_1,C_2). \end{aligned} \tag{2}$$

L'expérience d'enseignement d'ateliers de ce type confirme que cette troisième façon de voir les choses (appelée *modèle avec composante hypergéométrique* par la suite) est immanquablement avancée par des étudiants de culture biologique, issus d'un Master d'écologie par exemple, car c'est sous cette forme que le modèle standard de capture-marquage-recapture y est généralement présenté. Dans cette troisième vision, on a complètement symétrisé le rôle de C_1 et C_2 (on pourra vérifier l'invariance de la génération de C_{21} si on permute le rôle de C_1 et C_2 dans le tirage hypergéométrique). Pour un statisticien, elle est aussi l'occasion de souligner les problèmes délicats de dépendance conditionnelle et d'en discuter. En effet, il serait totalement déraisonnable d'utiliser de nouveau un modèle binomial $C_{21}|C_1,\pi_2 \sim dbin(\pi_2,C_1)$ pour la dernière opération ci-dessus, car une fois la seconde pêche C_2 réalisée, on ne peut plus *générer* C_{21} indépendamment de C_2, ne serait-ce que parce que $C_{21} \leq C_2$!

5. Estimation fréquentiste d'une taille de population inconnue, avec un coup de main de R

Comme illustré par la photo de la figure 2, le grand moment arrive : on procède enfin à l'expérience de CMR ! Le tableau 2 donne un exemple de résultats obtenus après deux pêches successives et lors d'une séance pendant laquelle les étudiants étaient repartis en trois groupes. Jusqu'à la fin de l'exercice, les étudiants de chaque groupe ignoreront l'effectif véritable de leur saladier. Mais, pour information, les voici : Groupe 1 : 210 ; Groupe 2 : 376 ; Groupe 3 : 244.

Tableau 2 – *Résultats d'une expérience de CMR à deux pêches avec trois groupes d'étudiants*

	Groupe 1	Groupe 2	Groupe 3
C_1	99	116	82
C_{20}	56	56	74
C_{21}	62	25	39

5.1. L'estimateur de Lincoln-Petersen

À ce stade, il convient d'interroger les étudiants sur leur meilleure estimée de l'effectif de leur population de poissons, non sans en avoir éventuellement profité pour rappeler brièvement la distinction entre *estimateur* et *estimation*. Il n'est alors pas rare que plusieurs d'entre eux supposent l'égalité des probabilités de capture ($\pi_1 = \pi_2$) et suggèrent ainsi que la proportion *connue* de poissons marqués lors de la seconde pêche soit une bonne estimation de la proportion *inconnue* de poissons marqués dans l'ensemble de la population (*i.e.* pris lors de la première pêche) : $\frac{C_{21}}{C_2} \approx \frac{C_1}{\nu}$. Poser l'égalité stricte des deux rapports permet de construire un estimateur ponctuel de l'effectif recherché : $\hat{\nu}_p = \frac{C_1 \times C_2}{C_{21}}$, appelé estimateur de Lincoln-Petersen (Chao et al., 2008). Cet estimateur est certes obtenu avec une approximation simple mais confondant probabilités et proportions. Hélas, cette confusion conceptuelle n'est par ailleurs pas nécessairement ressentie comme un raisonnement intuitif incomplet. Et hop, 188 poissons pour le premier groupe qui a bien envie de déclarer

mission accomplie et de ranger ses cahiers ! Comme beaucoup, sa bonne volonté s'arrête généralement lorsqu'on requiert une fourchette de crédibilité, par exemple les quartiles 25% et 75% autour de ce meilleur pari qu'il a calculé à partir de ses résultats expérimentaux, même en acceptant une réponse ni formalisée, ni justifiée, mais simplement issue d'une intuition ou d'une discussion au sein du groupe.

5.2. L'estimateur de Schnabel-Chapman

Soulignant le problème majeur posé par l'estimateur de Lincoln-Petersen quand $C_{21} = 0$, on présente alors un estimateur alternatif, appelé estimateur de Schnabel-Chapman et défini par : $\hat{\nu}_s = \frac{(C_1+1)\times(C_2+1)}{C_{21}+1} - 1$ (Amstrup et al., 2010). Viennent alors assez spontanément les deux questions suivantes : *Quel est le meilleur des deux estimateurs proposés ? Pourriez-vous proposer d'autres estimateurs ?* Le plus souvent, elles ne manquent pas de déstabiliser l'audience...

Pour soulager la tension qui commence à s'installer quand les étudiants réalisent qu'on les chatouille désagréablement sur leurs conceptions de l'incertitude, un travail empirique par simulations procure un secours appréciable ; par exemple, il est facile[10] d'écrire un programme en R qui simule 100000 répétitions de notre expérience CMR pour $\nu = 200, \pi_1 = \pi_2 = 0.1$, calcule les deux estimateurs, en trace la distribution empirique et évalue leurs caractéristiques (biais, variance et risque quadratique). La comparaison des deux estimateurs apparaît en figure 6 et, bien sûr, l'estimateur de Schnabel-Chapman se révèle non biaisé et de risque quadratique plus faible que celui de Lincoln-Petersen qui, de plus, peut-être non défini quand la probabilité de capture est suffisamment faible pour obtenir $C_{21} = 0$ avec une probabilité non négligeable. Encore faut-il bien faire comprendre que cette simulation s'effectue sous la loi d'échantillonnage, avec des paramètres qui ne sont plus inconnus mais fixés à une valeur que le modélisateur a donnée.

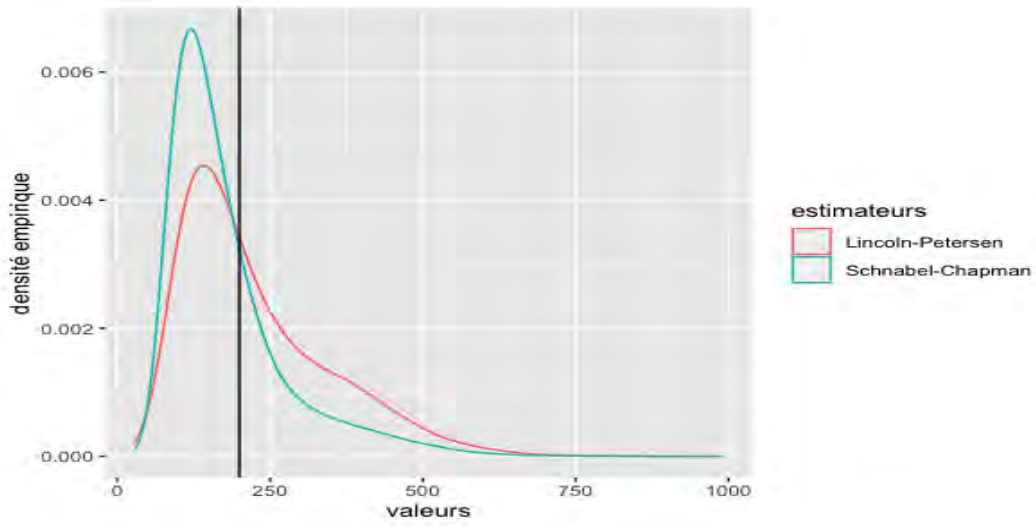


FIGURE 6 – *Comparaison de la répartition des estimateurs de Petersen-Lincoln et de Schnabel-Chapman pour une population* $\nu = 200$ *et une probabilité de capture fixée à* $\pi_1 = \pi_2 = 0.1$*. L'estimateur de Petersen-Lincoln n'est pas défini (*$C_{21} = 0$*) dans environ 13% des cas. L'espérance de l'estimateur de Petersen-Lincoln est ici évaluée à* 225 *quand il est défini, celle de Schnabel-Chapman à* 178*, avec pour risques quadratiques* $\mathrm{E}((\hat{\nu} - \nu)^2)$ *respectifs* 15090 *et* 10060*.*

Il est également possible d'entraîner les étudiants les plus férus de programmation informatique dans une étude par simulations un peu plus poussée dont l'objectif est de comparer l'évolution du

10. Voir le code des figures 6 et 7 en matériau supplémentaire de l'article.

biais relatif des estimateurs de Petersen-Lincoln et de Schnabel-Chapman en fonction de la taille de la population ν (allant par exemple de 100 à 1000 par pas de 50). On pourra par exemple, pour chaque valeur de ν, générer 20000 jeux de données de CMR en fixant la probabilité de pêche π à 0.30. La figure 7 illustre encore plus clairement que, contrairement à l'estimateur de Schnabel-Chapman, l'estimateur de Petersen-Lincoln est biaisé : il surestime la taille de la population ν et cette sur-estimation est d'autant plus marquée que la taille de la population est petite.

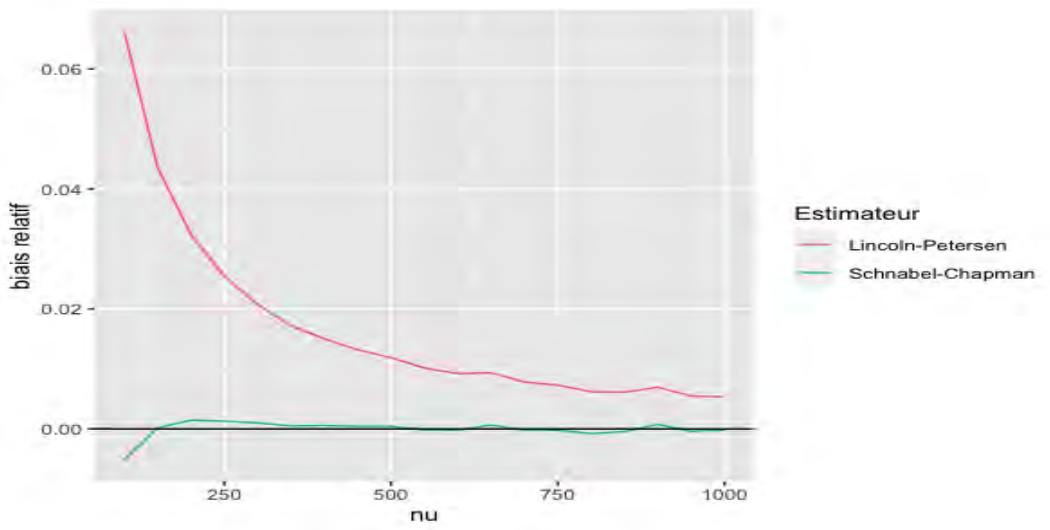


FIGURE 7 – *Évolution du biais (empirique) relatif des estimateurs de Petersen-Lincoln et de Schnabel-Chapman en fonction de la taille ν de la population, pour une probabilité de capture fixée à $\pi_1 = \pi_2 = 0.3$*

5.3. Simuler pour mieux comprendre ce que signifie la loi d'un estimateur

Continuons de procéder par simulation afin de comparer les propriétés des estimateurs proposés. Pour obtenir une évaluation empirique de l'écart-type des estimateurs de Lincoln-Petersen (quand celui-ci est défini) et Schnabel-Chapman, on peut effectuer de nombreux tirages dans la loi hypergéométrique (pour simuler des réalisations de C_{21}) dont on aura réglé les paramètres sur une première approximation de ν et π. De notre expérience, il ressort que nombre d'étudiants, même au niveau Master, n'ont pas perçu l'intérêt de la répétition simulée précédente pour évaluer le comportement d'un estimateur et, par conséquent, n'ont pas compris ce que représentait la loi d'échantillonnage et la cohabitation d'estimateurs multiples. Par ailleurs, force est de constater que les notions d'intervalle de confiance et de risque restent encore trop souvent non assimilées... Conseil au pédagogue : ne pas sortir de ses gonds, reprendre tranquillement les bases de la démarche statistique pour tenter de faire comprendre à son auditoire que se contenter d'une estimation ponctuelle, c'est être pressé d'avoir raison mais certainement pas une attitude scientifique responsable.

5.4. Calcul des moments de l'estimateur de Schnabel-Chapman

R possédant un générateur hypergéométrique *rhyper*, on pourrait bien sûr, d'après l'équation (2), simuler C_{21} sachant C_1 et C_2 pour faire une étude empirique des caractéristiques de l'estimateur de Schnabel-Chapman, mais pour les étudiants les plus férus de résultats théoriques (notamment ceux dans un parcours mathématique de niveau au moins M1), on peut les aider à entreprendre

vaillamment une analyse de portée plus générale. Pour en évaluer les propriétés théoriques, il faut alors s'appuyer sur les propriétés de la loi hypergéométrique.

Moyennes arithmetique et inverse Pour une variable aléatoire Y de loi hypergéométrique $(N+B, N, K)$, interprétée comme le nombre de boules noires obtenues après un tirage sans remise de K boules dans une urne contenant N boules noires et B boules blanches, on peut d'abord montrer les espérances mathématiques suivantes :

$$\begin{aligned} \mathrm{E}(Y) &= \frac{N}{N+B}K, \\ \mathrm{E}\left(\frac{1}{Y+1}\right) &= \frac{N+B+1}{(N+1)(K+1)}. \end{aligned} \tag{3}$$

En posant $K = C_2$, $N = C_1$, $B = \nu - C_1$ et $Y = C_{21}$, on démontre le caractère sans biais de l'estimateur de Schnabel-Chapman $\hat{\nu}_s$ dans les circonstances où C_1 et C_2 sont suffisamment grands pour que $C_1 + C_2 \geq \nu$. On trouve $\mathrm{E}(\hat{\nu}_s) = \nu$ par déconditionnement sur les variables aléatoires indépendantes C_1 et C_2.

Moyenne inverse d'ordre 2 Poursuivant l'étude des propriétés de l'hypergéométrique, on trouve dans Wittes (1972)[11] la preuve que la quantité

$$\begin{aligned} s^2 &= \frac{(C_1+1)(C_2+1)}{(C_{21}+1)^2}\frac{(C_1-C_{21})(C_2-C_{21})}{C_{21}+2} \\ &= \hat{\nu}_s^2 + 3\hat{\nu}_s + 2 - (C_1+1)(C_1+2)(C_2+1)(C_2+2)/(C_{21}+1)(C_{21}+2) \end{aligned}$$

est telle que

$$\mathrm{E}(s^2) = \mathrm{E}(\hat{\nu}_s^2) - \nu^2 = \mathrm{V}(\hat{\nu}_s),$$

c'est-à-dire que s^2 est un estimateur (sans biais) de la variance (conditionnelle à C_1 et C_2) de l'estimateur de Schnabel-Chapman.

Le tableau 3 récapitule moyenne et écart-type des estimateurs de Lincoln-Petersen (par simulation) et de Schnabel-Chapman (par la théorie) pour les données du tableau 2.

Tableau 3 – *Estimations et leurs écart-types des estimateurs (Petersen-Lincoln et Schnabel-Chapman) de l'effectif ν de poissons pour les 3 groupes d'étudiants (voir le code en matériau supplémentaire de l'article)*

Groupe	Petersen-Lincoln	Écart-type	Schnabel-Chapman	Écart-type
G1	188	10.6	188	9.9
G2	376	57.4	368	51.8
G3	238	23.2	236	21.4

5.5. L'estimateur du maximum de vraisemblance

Montons la barre ! Peu encourageant à la première lecture des équations de ce modèle, trouver le maximum de vraisemblance ne pose pourtant aucune difficulté numérique particulière pour les

11. Une erreur typographique s'est glissée dans Wittes (1972) : la quantité *inverse factorial moment* d'ordre k doit s'écrire $\mathrm{E}(\prod_{i=1}^{k}(n_{12}+i)^{-1})$ et non $\mathrm{E}(\prod_{i=1}^{k}(n_{12}+i)^{-2})$.

étudiants ayant assimilé leur cours d'optimisation. Supposons $\pi_1 = \pi_2$. En matériau supplémentaire de l'article, on montre que l'expression de $[C_1, C_{20}, C_{21}|\nu, \pi]$, vue comme une fonction de ν et π, est proportionnelle à

$$\frac{\nu!}{(\nu - C_1 - C_{20})!}\pi^{(C_1+C_2)}(1-\pi)^{2\nu-(C_1+C_2)}.$$

Maximiser cette fonction de vraisemblance revient à maximiser son logarithme, d'où : $\hat{\pi}_{mv} = \frac{C_1+C_2}{2\hat{\nu}_{mv}}$. Pour trouver $\hat{\nu}_{mv}$, on cherchera le maximum en ν de la fonction ci-dessous – dite de *vraisemblance profilée* – où on a substitué π par sa valeur optimale en fonction de ν :

$$\frac{\nu!}{(\nu - C_1 - C_{20})!} \times \left(\frac{C_1 + C_2}{2\nu}\right)^{(C_1+C_2)} \times \left(1 - \left(\frac{C_1 + C_2}{2\nu}\right)\right)^{2\nu-(C_1+C_2)}.$$

Pour maximiser cette expression, on pourra simplement parcourir toutes les valeurs (discrètes) possibles pour l'effectif de poissons ν.

Un intervalle de confiance pour cet estimateur peut s'obtenir de multiples manières, par exemple en ayant recours à l'approximation par un chi-deux de la déviance profilée (Cox et Hinkley, 1974; Casella et Berger, 2001). Mais rares furent les étudiants que nous avons rencontrés qui possédaient le niveau de culture suffisant en statistique mathématique asymptotique pour y parvenir.

Finalement, le tableau 4 résume les estimations de l'effectif ν de poissons et les intervalles de confiance asymptotiques à 95% obtenus pour chacun des 3 groupes d'étudiants pour les 3 estimateurs étudiés jusqu'à présent.

Tableau 4 – *Estimations et intervalles de confiance à 95% obtenus pour 3 estimateurs de l'effectif ν de poissons pour 3 groupes d'étudiants. Les suffixes utilisés sont "$_p$" pour Petersen-Lincoln, "$_s$" pour Schnabel-Chapman, "$_{mv}$" pour maximum de vraisemblance. Les intervalles de confiance à 95% calculés pour les estimateurs de Lincoln-Petersen et de Schnabel-Chapman supposent la normalité asymptotique de ces estimateurs. Celui du maximum de vraisemblance, lui dissymétrique, est obtenu grâce à l'approximation par un chi-deux de la vraisemblance profilée. Le code R figure en matériau supplémentaire de l'article.*

Groupe	$\hat{\nu}_{025.p}$	$\hat{\nu}_p$	$\hat{\nu}_{975.p}$	$\hat{\nu}_{025.s}$	$\hat{\nu}_s$	$\hat{\nu}_{975.s}$	$\hat{\nu}_{025.mv}$	$\hat{\nu}_{mv}$	$\hat{\nu}_{975.mv}$
G1	168	188	208	169	188	207	172	189	213
G2	258	376	494	267	368	469	297	385	534
G3	192	238	283	194	236	278	205	242	299

6. La piste bayésienne

Si l'on dispose de plus d'une journée pour mener cet atelier *Gommettes, haricots et saladier*, consacrer une séance à l'initiation au raisonnement bayésien (Robert, 2005; Parent et Bernier, 2007; Boreux et al., 2010; McElreath, 2020) représente un défi intéressant compte-tenu de deux de ses aspects de caractère typiquement bayésien : après tout, chaque groupe d'étudiants a préparé le matériel de ses voisins, et donc possède de l'information *a priori* sur l'ordre de grandeur du nombre possible de haricots dans un saladier et de l'efficacité de la cuillère à soupe comme instrument de pêche ! Et prendre un pari subjectif quant aux valeurs des variables du problème fait ici au moins autant sens qu'imaginer d'évaluer une fréquence d'occurrence lors d'une hypothétique répétition expérimentale.

6.1. Commodité de l'approche bayésienne

L'objectif d'un atelier CMR traité en bayésien sera essentiellement de faire saisir l'idée que de nombreuses difficultés techniques relatives à la mise en pratique de l'approche bayésienne sont aujourd'hui levées (Brooks, 2003). Pour une première initiation opérationnelle à l'inférence bayésienne, nous suggérons de s'appuyer sur les logiciels *clic-boutons* de la famille BUGS (Gilks et al., 1994; Lunn et al., 2000) et le langage déclaratif associé très proche de R. Pour un premier contact avec l'inférence bayésienne à partir d'algorithmes Monte-Carlo par Chaînes de Markov (MCMC), le package R baptisé « rjags »[12] qui permet d'appeler le logiciel Jags de Plummer (2015) a reçu notre faveur, car il fonctionne efficacement sous tous les systèmes d'exploitation de l'ordinateur : Windows, MacOs ou Linux. Sur nos données, pour le goupe 1, la figure 8 permet ainsi de visualiser[13] la loi *a posteriori* jointe de ν et π (pour le groupe d'étudiants 1) et de constater l'existence d'une corrélation *a posteriori* entre ces deux quantités.

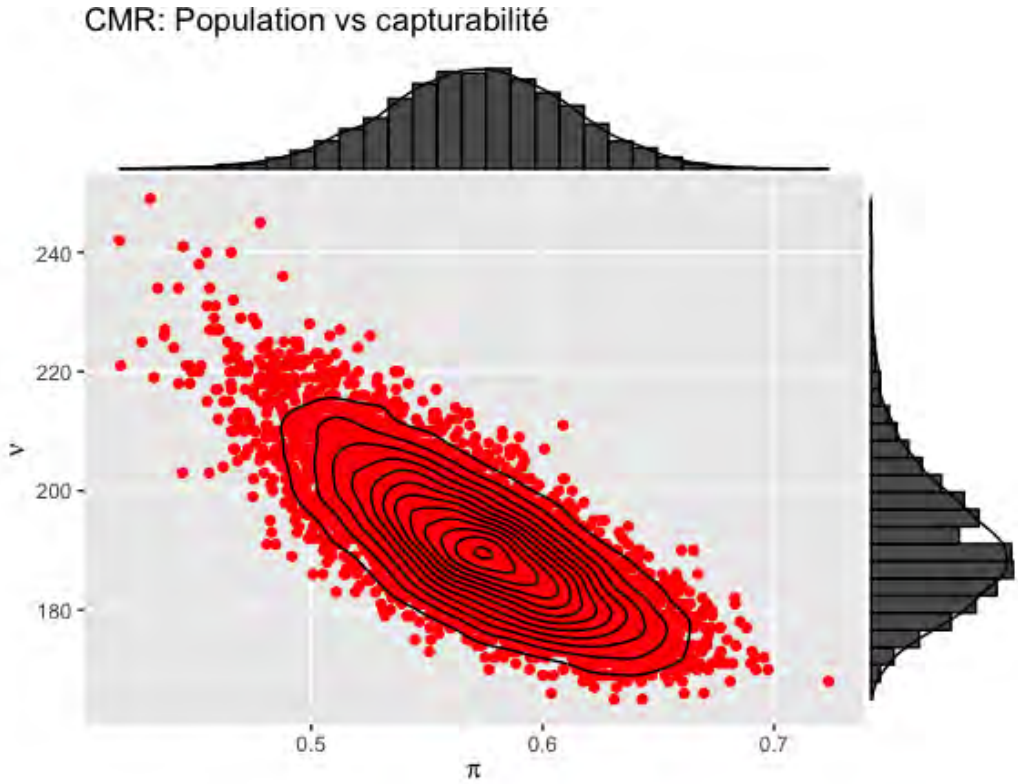


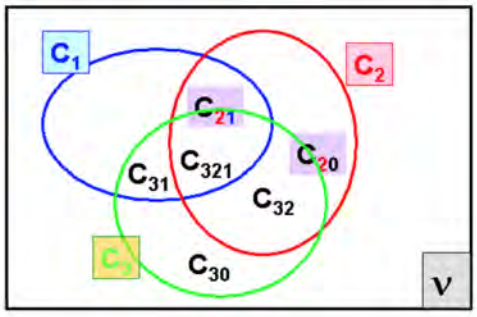
FIGURE 8 – *Loi a posteriori de la capturabilité π et de la taille de la population de poissons ν connaissant les données obtenues par le groupe 1 pour une loi a priori bêta(2,1) sur π et uniforme sur ν*

Les caractéristiques des lois marginales *a posteriori*, présentées dans le tableau 5 s'interprètent naturellement en terme de pari intuitif : pour le premier groupe, il y a 95 chances sur 100 que la taille de la population se situe entre 173 et 215 poissons (c'est-à-dire, *je suis prêt à parier à* 95 *contre* 5 *sur cette assertion*). Pour les groupes qui ont continué leur expérience de CMR en passant de deux à trois pêches successives avec remise (voir la figure 9), le problème inférentiel serait nettement plus compliqué à résoudre selon l'approche fréquentiste. Mais en bayésien, les résultats *a posteriori* obtenus avec seulement quelques lignes de programmation (voir le code BUGS en matériau supplémentaire) mettent facilement en évidence le gain de précision apporté par une troisième pêche sur la connaissance des inconnues : pour le groupe 1 par exemple, sachant les quatre informations

12. https ://cran.r-project.org/web/packages/rjags/index.html
13. L' obtention de cette figure et du tableau suivant est détaillée en matériau supplémentaire.

Tableau 5 – *Statistiques a priori et a posteriori relatives à l'effectif de poissons ν et à la capturabilité π pour le groupe 1. En gras figurent deux estimateurs bayésiens classiques (i.e. moyenne et médiane a posteriori) ainsi que les bornes de l'intervalle de crédibilité à 95% de ν.*

param	mean	sdt	q.025	q25	q50	q75	q.975
ν_prior	250	144	11	125	249	373	487
ν_post	**191**	11	**173**	183	**190**	197	**215**
π_prior	0.67	0.24	0.16	0.50	0.71	0.87	0.98
π_post	0.57	0.04	0.49	0.54	0.57	0.60	0.65



	Gp1	Gp3
C1	99	82
C20	56	74
C21	62	39
C30	41	40
C321	43	12
C31	15	14
C32	38	27

FIGURE 9 – *Diagramme de Venn (à gauche) et données recueillies par les groupes d'étudiants 1 et 3 (à droite) lors d'une expérience de CMR basée sur trois pêches successives*

supplémentaires apportées par la troisième pêche, la meilleure estimée du nombre d'individus se décale en effet de 191 à 215 tandis que l'écart-type *a posteriori* se réduit de 11 à 7.

6.2. Efficacité : l'intérêt pour un statisticien classique d'emprunter la piste bayésienne

Convenons pourtant que comparer des algorithmes (non des méthodes scientifiques) selon les résultats des calculs et leurs facilités de mise en œuvre ne veut rien dire en soi. Une raison plus subtile peut motiver un statisticien classique à emprunter *incognito* la piste bayésienne.

Pour les étudiants matheux de M1 ou M2, cette expérience CMR est aussi l'occasion de voir (ou de revoir) quelques éléments de la théorie de l'estimation. Une espérance *a posteriori*, comme celle évaluée pour ν au tableau 5 est une fonction (éventuellement non explicite) des données, il s'agit donc d'un estimateur au sens classique. Il est même d'ailleurs extrêmement intéressant pour un statisticien *classique* car c'est un estimateur non dominé (encore appelé efficace dans certains manuels de théorie de l'estimation). Par définition, un estimateur est efficace si on ne peut pas trouver un autre estimateur dont le risque quadratique soit uniformément meilleur sur tout le domaine de définition de ν. Or, le théorème de la classe complète de Wald (1947) établit une passerelle essentielle entre statistiques classique et bayésienne : il précise que tous les estimateurs efficaces (les seuls dignes d'intérêt pour le statisticien !) sont engendrés par la classe des estimateurs bayésiens et leurs limites.

6.3. Des propriétés structurelles pour modéliser ? de Finetti et l'échangeabilité

Du point de vue plus général de l'initiation à la modélisation probabiliste à partir d'une expérience de CMR, même la simple écriture binomiale séquentielle est déjà fort riche d'enseignements : les étudiants ont tendance à calquer un modèle d'urne avec tirages sans remise sur le mécanisme de pêche. Ce stéréotype n'est pas approprié ici : il s'agit plutôt d'une somme de comportements individuels Bernoulli *pris/pas pris* indépendants. Les conditions *iid* peuvent être ici postulées très raisonnablement : un poisson capturé n'alerte pas ses congénères, l'espace disponible ne modifie pas la capturabilité résiduelle. Par exemple, lorsqu'un dispositif par pêche électrique est utilisé, la technique de pêche fait que, dans le disque d'influence du dispositif électrique, tous les poissons, gros ou petits, sont attirés irrésistiblement vers l'anode tenue par le technicien. À un niveau plus avancé de l'enseignement, des raisons structurelles de symétrie plaident également pour un modèle binomial : il semblerait en effet déraisonnable de ne pas postuler l'invariance de la représentation probabiliste prédictive du système par toute permutation de l'ordre des individus capturés et, assumant cette propriété quelle que soit la taille de la population, le théorème de représentation de de Finetti (1937) impose alors une vraisemblance binomiale pour ces données individuelles binaires ! Ce théorème d'un grand auteur bayésien du siècle passé démontre également l'existence mathématique d'une quantité sur laquelle conditionner pour retourner à l'indépendance – nous avons ici proposé le couple (ν, π) – et de la loi qui doit lui être associée. Voilà que réapparaît la loi *a priori*, et une autre histoire de modélisation...

6.4. Interpréter la probabilité ?

À ce stade, hasarder un œcuménisme optimiste entre les postures fréquentiste et bayésienne s'avèrera toxique (Lecoutre, 1997) car les interprétations de la probabilité ne se recoupent pas du tout de façon *cohérente* au sens de Lindley (2013). Une p-value n'a rien d'une crédibilité bayésienne. La philosophie cognitive doit servir : expliquer pourquoi probabilité fréquentiste et crédibilité bayésienne ne sont pas la même chose participe à la rigueur intellectuelle et c'est un fait crucial pour la bonne formation d'un statisticien. Aussi se doit-on de prendre du temps pour répondre aux questions des étudiants concernant l'approche statistique bayésienne et en rappeler formellement les bases (Savage, 1954, 1971; Bernardo et Smith, 2009; Kadane, 2011; Lindley, 2013). D'ailleurs, comme ce paradigme semble d'emblée pour certains d'interprétation intuitive plus naturelle et immédiate (Collectif Biobayes, 2015; Lambert, 2018), la pilule de la cohérence mathématique que garantit le calcul des probabilités sera d'autant moins difficile à faire passer. Leur questionnement le plus inquiet portera sans doute sur la loi *a priori* : dans notre exemple, l'incorporation d'expertise probabilisée sur les valeurs des inconnues ν, π_1 et π_2. C'est simplement une question de modélisation (voir à ce propos le point de vue provocateur de Spiegelhalter et al. (2004) à la page 73 du chapitre 3 de leur ouvrage), mais la modélisation, ce n'est pas si facile ! Il faut dire qu'on peut discuter sans état d'âme du choix des lois *a priori* et réaliser une analyse de sensibilité, afin d'être conscient de l'impact potentiel de ces choix, au vu des données disponibles. Dans l'exemple de CMR, comme on ne fait pas de recensement, on est obligé de reconnaître que notre état de connaissance concernant la cible ν est incertain. Il y a donc un sens à représenter cette incertitude par une distribution de probabilité, et ce, même quand cet état de connaissance est très réduit. Par exemple, changer la loi uniforme sur la gamme des effectifs pour une loi uniforme sur le *logarithme* de ν, c'est considérer que la même chance est donnée à tous les ordres de grandeur possibles. Le choix d'une loi *a priori* bêta sur π de paramètres $a = 2$ et $b = 1$ serait un choix acceptable pour encoder, par exemple, le jugement d'un expert de la pêche électrique annonçant que son évaluation moyenne personnelle de l'efficacité π est de l'ordre de $2/3$ mais, sans grande confiance, puisqu'il n'est prêt à parier qu'une chance sur deux environ pour l'intervalle $[0.4, 0.8]$.

7. Épilogue

7.1. Perspectives

Plusieurs extensions plus spécifiques peuvent être apportées à cette expérience ludique et facile à réaliser en salle avec des gommettes, des haricots secs, une cuillère à soupe et un saladier. Citons, par exemple :

– Il est possible de poursuivre l'initiation aux algorithmes MCMC (Robert et Casella, 2013) : la faible dimension du problème permet de se lancer dans l'implémentation d'un algorithme de Gibbs simple et de comparer les résultats à ceux obtenus avec le logiciel Jags. Remarquons que les calculs peuvent se faire à la main en s'appuyant sur la conjugaison partielle (loi *a priori* bêta), mais il faut expliquer l'algorithme de Gibbs et éventuellement les techniques de Raoblackwellisation (Casella et Robert, 1996), notamment en ce qui concerne l'inconnue π. Les lois conditionnelles complètes sont ici explicites : loi bêta pour π et loi discrète pour ν (dérivées en matériau supplémentaire). La littérature sur les échantillonneurs par algorithmes markoviens est en constante évolution (Lunn et al., 2009) ; d'autres outils d'échantillonnage, par exemple du type STAN (Gelman et al., 2015) ou NIMBLE (de Valpine et al., 2017), peuvent également être testés à partir de notre expérience jouet.

– Certaines perspectives intéresseront davantage les écologues que les probabilistes. Il n'est guère difficile d'adapter le matériel de notre expérience de CMR pour faire une introduction aux techniques de capture avec enlèvements successifs, un autre moyen d'usage courant en écologie pour évaluer la taille d'une population (Rivot et al., 2008) ou pour s'initier à la représentation d'un système dynamique à état discret en adoptant des règles de mortalité et de naissance sur la population (King et al., 2009). Enfin les méthodes de CMR font l'objet de constructions hiérarchiques fructueuses (Rivot et Prévost, 2002) pour représenter les ressemblances entre années, sites, etc., et en tirer profit pour une inférence plus riche d'informations.

7.2. Conclusions

Dans la France des années 1650, le Chevalier de Méré avait un entêtant problème de jeu :

– Lancer un dé équilibré au maximum quatre fois, et gagner si vous obteniez un six ;

– Lancer deux dés au maximum vingt-quatre fois pour obtenir un double-six.

Quel était le meilleur pari ?

Reprenant la solution erronée que le Chevalier de Méré en avait présentée, Blaise Pascal et Pierre de Fermat planchèrent sur le problème. L'histoire des sciences retient que c'est ainsi qu'ensemble, ils développèrent les premiers éléments de la théorie des probabilités...

Finalement, l'expérience que nous proposons ici – ludique et facile à effectuer en salle avec des gommettes, des haricots secs, une cuillère à soupe et un saladier – n'est peut-être qu'un simple retour aux sources, avec le jeu, ses paris sur les résultats possibles et l'observation répétée de données expérimentales. A partir de données réelles, collectées par les étudiants eux–mêmes, elle permet de développer, dès les premières séances d'un cours de statistique inférentielle, de nombreux points-clés du raisonnement probabiliste – qu'il soit fréquentiste ou bayésien – indispensables au statisticien-modélisateur. Gageons qu'en l'abordant de façon délibérément simple et empirique, elle suscitera néanmoins la réflexion et mobilisera la faculté d'abstraction des étudiants face aux questions liées à la quantification des incertitudes.

Matériau supplémentaire

Les auteurs sont convaincus que la reproductibilité totale est la norme minimale pour juger des travaux scientifiques. Un fichier *html* est disponible pour vérifier et reproduire tous les chiffres et résultats de cet article à la page web de la revue.

Remerciements

Les auteurs remercient Jacques Bernier pour les nombreuses discussions, souvent vigoureuses mais toujours constructives, concernant la cohérence et la rationalité du discours statistique.

Références

Amstrup, S., T. McDonald, et B. Manly (2010), *Handbook of Capture-Recapture Analysis*, Princeton University Press, URL http://books.google.fr/books?id=hOJxGNERUKgC.

Annis, D. H. (2005), «Rethinking the paper helicopter : Combining statistical and engineering knowledge», *The American Statistician*, vol. 59, nº 4, pp. 320–326.

Azaïs, J.-M. (2004), «Illustration de la méthode des plans d'expériences sur la comparaison de boissons au cola», *Journal de la société française de statistique*, vol. 145, nº 4, pp. 69–78.

Bernardo, J. M. et A. F. Smith (2009), *Bayesian theory*, vol. 405, John Wiley & Sons.

Boreux, J.-J., E. Parent, J. Bernier, et J. Bernier (2010), *Pratique du calcul bayésien*, vol. 118, Springer.

Box, G. E. (1992), «Teaching engineers experimental design with a paper helicopter», *Quality Engineering*, vol. 4, nº 3.

Brooks, S. P. (2003), «Bayesian computation : a statistical revolution», *Philosophical Transactions of the Royal Society of London. Series A : Mathematical, Physical and Engineering Sciences*, vol. 361, nº 1813, pp. 2681–2697.

Casella, G. et R. L. Berger (2001), *Statistical inference*, Duxbury/Thomson Learning.

Casella, G. et C. P. Robert (1996), «Rao-blackwellisation of sampling schemes», *Biometrika*, vol. 83, nº 1, pp. 81–94.

Chao, A., H.-Y. Pan, et S.-C. Chiang (2008), «The petersen–lincoln estimator and its extension to estimate the size of a shared population», *Biometrical Journal : Journal of Mathematical Methods in Biosciences*, vol. 50, nº 6, pp. 957–970.

Collectif Biobayes (2015), *Initiation à la statistique bayésienne : bases théoriques et applications en alimentation, environnement, épidémiologie et génétique*, Ellipses.

Collett, D. (2002), *Modelling binary data*, CRC press.

Cox, D. R. et D. V. Hinkley (1974), *Theoretical Statistics*, Chapman and Hall, London.

de Finetti, B. (1937), «La prévision : ses lois logiques, ses sources subjectives», in «Annales de l'institut Henri Poincaré», vol. 7, pp. 1–68.

de Valpine, P., D. Turek, C. J. Paciorek, C. Anderson-Bergman, D. T. Lang, et R. Bodik (2017), «Programming with models : writing statistical algorithms for general model structures with nimble», *Journal of Computational and Graphical Statistics*, vol. 26, nº 2, pp. 403–413.

Dudley, B. (1983), «A practical study of the capture/recapture method of estimating population size», *Teaching Statistics*, vol. 5, nº 3, pp. 66–70.

Gelfand, A. E. et A. F. Smith (1990), «Sampling-based approaches to calculating marginal densities», *Journal of the American Statistical Association*, vol. 85, nº 410, pp. 398–409.

Gelman, A., D. Lee, et J. Guo (2015), «Stan : A probabilistic programming language for Bayesian inference and optimization», *Journal of Educational and Behavioral Statistics*, vol. 40, nº 5, pp. 530–543.

Gelman, A. et D. Nolan (2017), *Teaching statistics : A bag of tricks*, Oxford University Press.

Gilks, W. R., A. Thomas, et D. J. Spiegelhalter (1994), «A language and program for complex Bayesian modelling», *Journal of the Royal Statistical Society : Series D (The Statistician)*, vol. 43, nº 1, pp. 169–177.

Kadane, J. B. (2011), *Principles of uncertainty*, CRC Press.

King, R., B. Morgan, O. Gimenez, et S. Brooks (2009), *Bayesian analysis for population ecology*, CRC press.

Lambert, B. (2018), *A Student's Guide to Bayesian Statistics*, Sage.

Lecoutre, B. (1997), «C'est bon à savoir», *Et si vous étiez un bayésien qui s' ignore. Modulad*, vol. 18, pp. 81–87.

Leyland, A., M. Barnard, et N. McKeganey (1993), «The use of capture-recapture methodology to estimate and describe covert populations : An application to female street-working prostitution in Glasgow», *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, vol. 38, nº 1, pp. 52–73.

Lindley, D. V. (2013), *Understanding uncertainty*, John Wiley & Sons.

Lunn, D., D. Spiegelhalter, A. Thomas, et N. Best (2009), «The bugs project : Evolution, critique and future directions», *Statistics in medicine*, vol. 28, nº 25, pp. 3049–3067.

Lunn, D. J., A. Thomas, N. Best, et D. Spiegelhalter (2000), «Winbugs-a Bayesian modelling framework : concepts, structure, and extensibility», *Statistics and computing*, vol. 10, nº 4, pp. 325–337.

Marin, J.-M. et C. Robert (2007), *Bayesian core : a practical approach to computational Bayesian statistics*, Springer Science & Business Media.

McCrea, R. S. et B. J. Morgan (2014), *Analysis of capture-recapture data*, CRC Press.

McElreath, R. (2020), *Statistical rethinking : A Bayesian course with examples in R and Stan*, CRC press.

Parent, E. et J. Bernier (2007), *Le raisonnement bayésien : modélisation et inférence*, Springer Science & Business Media.

Plummer, M. (2015), «Jags version 4.0. 0 user manual», *Lyon. Available online at : http ://sourceforge.net/projects/mcmc-jags*.

Rivot, E. et E. Prévost (2002), «Hierarchical Bayesian analysis of capture mark recapture data», *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 59, nº 11, pp. 1768–1784.

Rivot, E., E. Prévost, A. Cuzol, J.-L. Baglinière, et E. Parent (2008), «Hierarchical Bayesian modelling with habitat and time covariates for estimating riverine fish population size by successive removal method», *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 65, nº 1, pp. 117–133.

Robert, C. (2005), *Le choix bayésien : Principes et pratique*, Springer Science & Business Media.

Robert, C. et G. Casella (2013), *Monte Carlo statistical methods*, Springer Science & Business Media.

Royle, J. A., R. B. Chandler, R. Sollmann, et B. Gardner (2013), *Spatial capture-recapture*, Academic Press.

Savage, L. J. (1954), *The foundations of statistics*, Courier Corporation.

Savage, L. J. (1971), «Elicitation of personal probabilities and expectations», *Journal of the American Statistical Association*, vol. 66, nº 336, pp. 783–801.

Seber, G. A. F. et al. (1982), *The estimation of animal abundance and related parameters*, vol. 8, Blackburn press Caldwell, New Jersey.

Spiegelhalter, D. J., K. R. Abrams, et J. P. Myles (2004), *Bayesian approaches to clinical trials and health-care evaluation*, vol. 13, John Wiley & Sons.

Spiegelhalter, D. J., A. P. Dawid, S. L. Lauritzen, et R. G. Cowell (1993), «Bayesian analysis in expert systems», *Statistical science*, pp. 219–247.

Tibshirani, R. J., A. Price, et J. Taylor (2011), «A statistician plays darts», *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, vol. 174, nº 1, pp. 213–226.

Wald, A. (1947), «An essentially complete class of admissible decision functions», *The Annals of Mathematical Statistics*, pp. 549–555.

Wikle, C. K., L. M. Berliner, et N. Cressie (1998), «Hierarchical Bayesian space-time models», *Environmental and Ecological Statistics*, vol. 5, nº 2, pp. 117–154.

Wittes, J. T. (1972), «Note : On the bias and estimated variance of Chapman's two-sample capture-recapture population estimate», *Biometrics*, pp. 592–597.

EMOS – Towards a unified training of European public statisticians

Annika NÄSLUND[1]
Former Chair of the EMOS Board
Head of Unit "Planning and Evaluation; Statistical training" at Eurostat, European Commission

TITLE
EMOS – Vers une formation unifiée des statisticiens publics européens

RÉSUMÉ
L'article décrit comment est née l'idée de créer un Master européen en statistiques officielles (European Master in Official Statistics, EMOS), comment le concept a été développé et mis en œuvre, ainsi que certains des défis qui restent à relever pour l'avenir.

Mots-clés **:** *master, statistiques officielles, Europe, Eurostat, Système statistique européen.*

ABSTRACT
The article describes how the idea to launch a European Master in Official Statistics (EMOS) was born, how the concept was developed and implemented, as well as some of the challenges that remain for the future.

Keywords**:** *master, official statistics, European, Eurostat, European Statistical System.*

1. Annika.Naslund-Fogelberg@ec.europa.eu

1. Introduction

In order to understand the context in which the European Master in Official Statistics was launched, it is important to first briefly explain the roles and responsibilities of two of the main actors behind EMOS, Eurostat and the European Statistical System.

Eurostat is a Directorate-General of the European Commission and the statistical office of the European Union. Its mission is to provide high-quality statistics for Europe. As explained in the Regulation on European Statistics[2] and in the Commission Decision on Eurostat[3], Eurostat ensures the development, production and dissemination of European statistics according to established rules and statistical principles, notably those laid down in the European statistics Code of Practice[4]. Moreover, Eurostat coordinates the statistical activities of the institutions and bodies of the Union, in particular with a view to ensuring consistency and quality of the data and minimising reporting burden.

The European Statistical System is the partnership between the statistical authority of the Union, which is the European Commission (Eurostat), and the National Statistical Institutes (NSIs) and other national authorities responsible in each Member State for the development, production and dissemination of European statistics. This partnership also includes the European Economic Area and European Free Trade Association countries.

The European Statistical System functions as a network in which Eurostat's role is to lead the way in the harmonisation of statistics, through the harmonisation of methodologies, concepts and classifications, in close cooperation with the national statistical authorities, which collect the data in the agreed format and transmit them to Eurostat. Eurostat consolidates the data, produces European aggregates and publishes high-quality comparable European statistics. As EU cooperation covers more and more EU policy areas, the harmonisation of statistics has over the years been extended to nearly all statistical fields.

2. The European Master in Official Statistics (EMOS)

2.1 What is EMOS?

EMOS is essentially a label that is awarded by the European Statistical System Committee to university Master programmes that fulfil the EMOS eligibility and selection criteria (see Section 2.3.3) of higher level education in the field of official statistics. For the moment, there are 32 EMOS-labelled Master programmes in 19 European countries. The label is valid for four years and renewable under condition that the criteria are still fulfilled. An important part of the selection criteria is that the university applying for the label should have an established cooperation with the national statistical institute or another authority producing official statistics. This ensures that the gap between theory and practice is bridged in a good way. Indeed, one of the requirements is that the Master theses and the internships, which are compulsory, are carried out by students in the field of official statistics in close cooperation with a producer of official statistics.

2.2 Why was EMOS created?

The idea to create a European Master in Official Statistics was born against the backdrop of the very rapid evolution of the competences required to manage the whole stream of processes necessary to design, produce and disseminate official statistics and the need to strengthen the value of statistics in modern societies.

2. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02009R0223-20150608
3. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2012:251:TOC
4. https://ec.europa.eu/eurostat/documents/4031688/8971242/KS-02-18-142-EN-N.pdf/e7f85f07-91db-4312-8118-f729c75878c7

As a starting position, the overall aim of EMOS was to provide postgraduate training in official statistics for professionals who are able to work with official statistics at different levels and to reinforce the collaboration between universities and producers of official statistics.

Early discussions on the establishment of a European Master in Official Statistics originate from parallel initiatives of, on the one side, the French national statistical institute and GENES[5] and a few other national statistical institutes and, on the other, Eurostat. A common framework for training of "official statisticians" had already been discussed in the past, but the first real analysis was started at the "*Workshop on a European Master in Official Statistics*", jointly organised by Eurostat and the University of Southampton in 2010, where national statistical institutes and academia from over 20 European countries gathered to brainstorm about this topic.

One of the conclusions was to create a group of volunteers to further reflect on how to take some of the ideas from the workshop forward. The group, which was chaired by the Italian national statistical institute, included participants from different national statistical institutes, universities and Eurostat. The work of the group resulted in a paper outlining a first set of ideas on how to establish the necessary framework, which was presented to the European Statistical System Committee in February 2012.

The following main objectives were defined:

- Reinforce the network of professional statisticians at international level.
- Strengthen the cooperation between academia and the ESS, e.g. on research topics of relevance for official statistics.
- Ensure a wider offer of higher education in the area of official statistics in Europe and curricula adapted to the changing needs of statistical authorities.
- Develop a shared vision on methodology, organisation and management of the production of European statistics, by involving both academia and the European Statistical System as teaching parties in EMOS.
- Meet training and recruitment needs by constituting a future recruitment pool of highly educated professional statisticians for the European Statistical System.

2.3 How was the EMOS concept developed?

Once Eurostat and the European Statistical System Committee had agreed to go ahead with EMOS, a feasibility study was launched in order to assess both feasibility and interest among the main stakeholders. It was also necessary to come up with a governance structure and to define the conditions under which Master programmes could join the EMOS network.

2.3.1 Feasibility study

The main objectives of the feasibility study, which was carried out by a consortium of ICON-INSTITUT Consulting Group and GENES, were to:

- Provide an inventory of Master programmes in statistics and their providers in the selected countries.
- Analyse the existing and potential Master programmes in official statistics and assess if they were suitable for joining a future European network of Masters in official statistics.
- Assess the interest of stakeholders – in particular universities and NSIs.
- Analyse advantages/disadvantages, the cost-benefit, labelling mechanisms as well as the role and implications for NSIs and Eurostat.
- Propose a roadmap for the implementation of EMOS.

5. Le Groupe des Écoles Nationales d'Économie et Statistique (GENES) is a public institution of higher education and research attached to the Ministry of Economy and Finance, for which INSEE provides human support and technical oversight. On the reverse side, it provides the initial training of INSEE's executives.

The feasibility study started in December 2012 and took one year to complete. It was divided into three phases.

In the first phase, a quantitative online survey was launched among universities offering Masters in statistics in 39 different European countries. The survey addressed issues such as the Master programme's alignment with the Bologna process and the European Credit and Transfer System (ECTS), the duration of the Master, the existence of a scientific committee, accreditations rules, the fields of the Master and existing cooperation with the national statistical institute. Out of the over 700 universities addressed by the survey, 151 valid replies were received.

In the second phase, 41 universities and 14 national statistical institutes in 14 different countries were interviewed to assess their willingness to participate in the EMOS network, the structure of the Master and its international relevance, cooperation between the university and the National Statistical Institute and the availability of resources (e.g. communication tools, computers, courses in English).

The results of the two surveys showed that:

- The Bologna Process was adhered to in almost all countries.
- Almost all Master programmes had a scientific committee and an accreditation system in place.
- Two thirds of the Masters had connections with the NSI.
- Almost all Masters had a traditional educational structure with lecturers and academic staff.
- At least 25 programmes would be ready to implement EMOS based on the criteria proposed by the feasibility study with only minor changes in their curricula. Other universities would be willing to join, but their Masters were not fulfilling all the criteria yet.
- A majority of participating countries were interested in EMOS, as long as there would be no impact on the core programme of their existing Master, as this could imply accreditation issues.
- Two categories of Master programmes could be distinguished: the professional Masters and those more focused on research.
- Financing was a concern for some universities, notably for staff and teacher exchanges, as well as course material in English.
- In some countries, issues such as political and economic specificities, existing regulations of higher educational systems and English as the main teaching language, would need to be carefully considered.

In a third phase, an analysis of advantages and disadvantages, cost-benefits, labelling mechanisms as well as roles and implications of stakeholders were analysed and a concept for a possible European Master in Official Statistics prepared.

The most important recommendations from the feasibility study were the following:

- Given the different national accreditation procedures for new Master programmes and the complexity of such procedures in most countries, it was recommended to implement EMOS as a label for already existing Master programmes, which comply with a set of EMOS criteria. In other words, rather than creating a new Master from scratch without certainty about its attractiveness among students, the aim would be to build on already existing programmes, by adding an EMOS module ensuring a lowest common denominator of statistical knowledge and skills.
- To avoid accreditation issues, EMOS should have no major impact on the core parts of the existing Master programmes.

A Group of Experts (composed of universities, NSIs and Eurostat), was asked to accompany the feasibility study and to develop a proposal for a concrete draft curriculum, rules and conditions for Master programmes to join the network (labelling), as well as a governance structure.

2.3.2 Governance

The governance model that was suggested involved the following main actors: universities, a Board, the European Statistical System Committee and Eurostat. The structure and the actors' respective roles are briefly explained below.

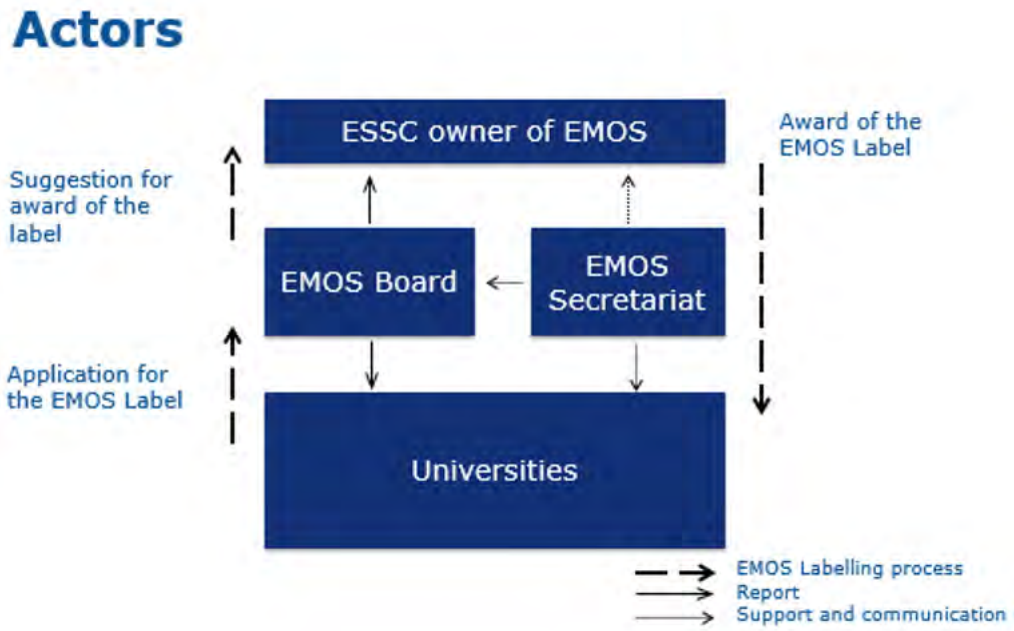


Figure 1 – *EMOS Governance model*

(i) The European Statistical System Committee

The European Statistical System Committee[6] is the highest level committee within the European Statistical System. It is composed of all Directors General of the national statistical institutes in the EU and EFTA countries, with a number of observers, notably the Directorate General for Statistics of the European Central Bank.

In the EMOS governance model (see Figure 1), the European Statistical System Committee is the owner of the EMOS label and as such the authority that awards the label to Master programmes fulfilling the EMOS requirements. It may consult the Board on any matter relating to education and training in official statistics and it also nominates the members of the EMOS Board.

(ii) Universities

Universities that are interested and fulfil the criteria may apply for the EMOS label through calls for applications organised by the EMOS Secretariat. The label is awarded for four years, after which the university can re-apply for the label if they still fulfil the requirements. In order to ensure feedback on quality, attractiveness and potential changes to the curricula of EMOS-labelled programmes, universities are also required to provide an annual report to the EMOS Board.

6. https://ec.europa.eu/eurostat/web/european-statistical-system/ess-governance-bodies/essc

(iii) EMOS Board

The mandate of the EMOS Board was approved by the European Statistical System Committee in September 2014. Its main role is to advise the European Statistical System Committee on matters relating to EMOS and report back to it. More specifically, the Board:

- assists the European Statistical System Committee with regard to the development of EMOS, in particular the award of the EMOS label to Master programmes fulfilling the EMOS requirements.
- contributes to quality monitoring of the EMOS Master programmes in order to ensure that the required standards are achieved and maintained through evaluation of new applications and extensions of the EMOS label.
- assists the European Statistical System Committee in meeting learning and development needs in the European Statistical System and advises on questions relating to higher education and training in official statistics, as well as further contributes to an exchange of experiences and good practices in the area of EMOS.

The chair of the Board may also advise the European Statistical System Committee to consult the Board on a specific question.

The first EMOS Board was appointed by the European Statistical System Committee in November 2014 for a period of three years. The Board is chaired by Eurostat and has 13 other members: six from universities, five from national statistical institutes, one from a national central bank and one from the European Statistical Advisory Committee[7].

(iv) EMOS secretariat

The EMOS secretariat is organised and staffed by Eurostat. The secretariat is responsible for all administrative and organisational matters. It organises the meetings and provides support to the EMOS Board, in particular in the EMOS label application process, and prepares the EMOS Board's recommendations and reports to the European Statistical System Committee. The secretariat also ensures communication and promotion of EMOS through a dedicated website and regular communication with the EMOS network and other interested parties.

2.3.3 Labelling

Given the national differences regarding the accreditation of new Master programmes, which became evident in the feasibility study, it was decided to implement EMOS as a label for existing Master programmes fulfilling certain requirements. A label can be considered as a supplement to accreditation and offers the necessary flexibility for universities wishing to introduce EMOS, without implications on national accreditation practices. It was agreed that the EMOS label must stand for excellence in European statistics and apply quality standards recognised by all concerned stakeholders.

Two sets of criteria were established, eligibility and selection criteria, as explained below.

(i) Eligibility criteria

The three eligibility criteria are:

- The Master programme should be an already accredited Master programme in an EU Member State, an EU candidate country or an EFTA country.
- The programme should be in line with the Bologna process.
- The programme should be in line with the European Credit Transfer and Accumulation System (ECTS).

7. https://ec.europa.eu/eurostat/web/european-statistical-system/ess-governance-bodies/esac

(ii) Selection criteria

As part of the selection criteria, the Master programme should have a workload equivalent to at least 90 ECTS, whereof:

- at least 50 ECTS should be devoted to the EMOS learning outcomes (see Figure 2 below);
- a Master thesis on a topic of official statistics, worth at least 20 ECTS;
- an internship in official statistics worth at least 10 ECTS, or, of at least 6 weeks' duration in cases where the university's rules do not allow awarding ECTS to the internship.

In addition, any university applying for the EMOS label should demonstrate that it has an active cooperation with the respective national statistical institute or other producers of official statistics, which is particularly important for the purposes of teaching, supporting Master theses and hosting internships. Moreover, the application should demonstrate that the programme includes lecturers with a sufficiently solid background in official statistics. Applicants also have to commit to starting their EMOS-labelled Masters at the latest one year after having been labelled.

The five EMOS learning outcomes have been designed to clarify what an EMOS student is expected to have learnt when graduating (see Figure 2).

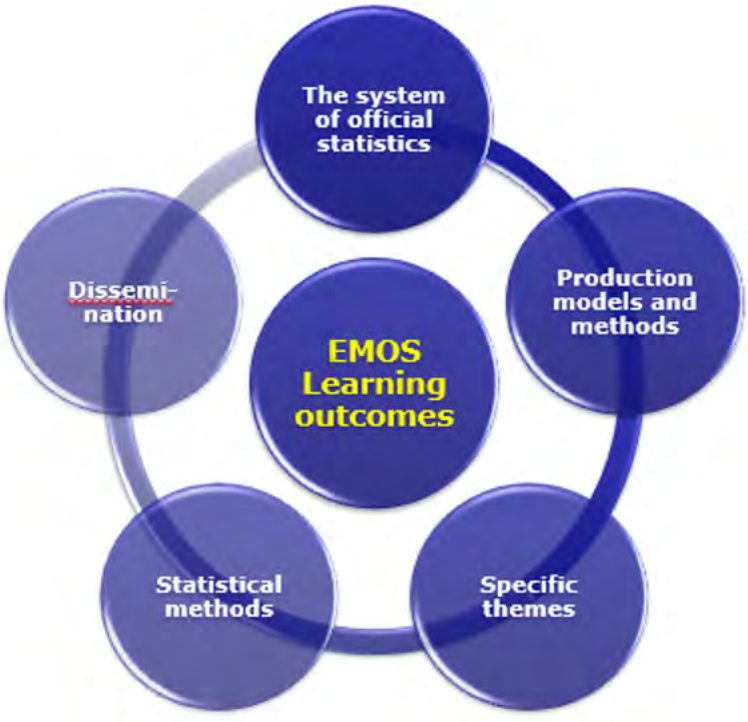


Figure 2 – *EMOS Learning outcomes*

1. The system of official statistics (e.g. relevance of official statistics, organisation and role of the European statistical system and other official data producers, main national, European and international institutions and data sources, principles of the European Statistics Code of Practice).
2. Production models and methods used for official statistics (including the use of multisource statistics and related quality dimensions).
3. Specific themes (e.g. economy, finance, population and social conditions, international trade, environment, energy, science and technology) and related methodological issues.
4. Statistical methods (e.g. sampling and estimation techniques, index theory, econometrics, statistical programmes).
5. Dissemination (communicating statistics to different audiences, use of different tools for

disseminating data and metadata, confidentiality and disclosure control).

3. EMOS Activities

At the start of EMOS in 2014, a lot of work had to be devoted to developing the rules and procedures for EMOS, raising awareness about the project, putting in place the governance structure and its reporting lines. Subsequently, three rounds of labelling were implemented and a number of activities to foster cooperation and sharing of experiences within the EMOS network, as briefly described below.

3.1 Labelling

The main role of the EMOS Board and its most time-consuming task has until now been the labelling process, involving amongst others the evaluation of the applications received for the EMOS label and the preparation of recommendations for the European Statistical System Committee, as owner of the EMOS label.

As mentioned above, the label is valid for four years, meaning that universities that wish to maintain their label need to re-apply for it and prove that they still fulfil the EMOS criteria. Assessing these applications is also part of the EMOS Board's activities and follows the same procedure as for the initial labelling.

In addition, the EMOS Board monitors the quality of the labelled programmes through the annual reporting by universities.

As shown in Figure 3, there are now 32 EMOS labelled Master programmes in 19 countries.

Figure 3 – *Map of EMOS labelled universities*

3.2 Webinars

Series of online lectures (webinars) on topics of official statistics are organised on a regular basis with speakers from cooperating universities, national statistical offices and Eurostat. The webinars are one-hour long and includes the possibility for students across Europe to interact with the lecturers. The webinars are freely accessible.

3.3 Workshops

Approximately, once per year an EMOS workshop is organised for representatives from EMOS-labelled Master programmes, national statistical institutes and the EMOS Board.

Participants discuss topics of common interest and can learn from each other's experiences. Examples of past topics include: Value and benefits of EMOS, the European dimension of EMOS, the role of NSIs and NCBs, teaching material for official statistics, Master theses and internships, audio-visual and online tools, best practices in teaching, official statistics and research. Last, but not least, an important part of the workshop is networking.

Illustration 1 – *EMOS Workshop in Ljubljana, March 2018 (photo: Francine Kessler)*

3.4 European Statistical Week and EMOS Open Days

The European dimension of EMOS becomes very concrete during two particular events organised by Eurostat on an annual basis, the European Statistical Week and the Open Days.

The European Statistical Week is a one-week study visit to Eurostat included in the European Statistical Training Programme[8]. The study visit, which is primarily targeted at young statisticians already working in a national statistical institute, also sees the participation of 10-15 EMOS students every year.

During this week, two and a half days are devoted to lectures where the students learn more about Eurostat and the European Institutions, the European Statistical System and statistical topics. Two days of job shadowing in a Eurostat unit are also included. The study visit provides participants with a concrete picture of what it is like to work at Eurostat and in the European Statistical System, which is usually highly appreciated by the participants. Moreover, useful contacts are established for future exchanges between colleagues from different national statistical institutes across Europe, as well as EMOS students.

For those in charge of EMOS-labelled Master programmes at the different universities, EMOS

7. https://ec.europa.eu/eurostat/fr/web/european-statistical-system/training-programme-estp

Open Days are organised. Like in the study visits mentioned above, participants in the Open Days learn more about Eurostat and the European Statistical System and some of the most important ongoing statistical work for the moment. They also have the opportunity to discuss EMOS related matters with Eurostat's senior management and get to meet and exchange with peers in other countries.

3.5 Master thesis competition

For the first time in 2018, an EMOS Master thesis competition was organised. The purpose of the competition was to reward outstanding Master theses amongst the network of, then, 23 EMOS labelled programmes. Its aim was to highlight official statistics as a research topic and put forward young talents with innovative contributions. The submitted Master theses were evaluated by the EMOS Board based on a set of clearly defined criteria and as a result, five Master theses were selected. The winners' prize was an invitation to present their theses at the New Techniques and Technologies for Statistics (NTTS) international biennial scientific conference[9], organised by Eurostat, in March 2019 in Brussels, with the expenses covered by Eurostat. It was the first time the students presented their work to an international audience and at such a big conference. Based on the positive feedback received from the audience, a dedicated session on EMOS will be proposed also for the next NTTS conference in 2021.

4. Challenges for the future

In 2017, the EMOS Secretariat carried out a first assessment of the implementation of EMOS. Feedback was collected from the main stakeholders, i.e. students (both following an EMOS-labelled Master programme and other Masters) universities (with and without EMOS-labelled Master programmes), producers of European statistics and the EMOS Board. After only two rounds of labelling, there was at the time not yet a lot of experience with EMOS, nor many EMOS graduates. Nevertheless, most of the findings were confirmed in subsequent annual reports submitted by universities.

While the feedback showed that EMOS was on the right track, had increased cooperation between universities and producers of official statistics and had successfully managed to link theory and practice, much thanks to the compulsory internships in official statistics, it also revealed a number of challenges requiring further work:

- low number of students enrolled in EMOS programmes (many of which were already working in a national statistical institute)
- the European dimension, including activities at European level, such as student and teacher cross-border exchanges and similar
- curricula would need to be modernised to better cover topics linked to the use of new and multiple data sources, big data, etc.
- commitment from the producers of official statistics

Based on the evaluation, the European Statistical Committee asked the EMOS Board to follow up on a set of recommendations linked to these challenges. With 23 EMOS-labelled Master programmes at the time (2017), it was also agreed that there was still scope to extend the network, so the go-ahead was given for the launch a third call for applications in 2018.

As the EMOS network of labelled Master programmes has now reached as many as 32 programmes in 19 different countries, it was agreed that focus should no longer be on expanding the network, but more on integration. Nevertheless, interested universities may still join the EMOS network through the permanently open call for applications with annual cut-off dates for submission of applications (this year 31 December 2020)[10].

9. https://ec.europa.eu/eurostat/cros/content/emos-events_en
10. https://ec.europa.eu/eurostat/cros/content/2020-emos-call-applications_en

The main challenges ahead will therefore be to attract students to EMOS and to strengthen the European dimension, e.g. by developing cross-border collaboration within the EMOS network through student and teacher exchanges and internships. Here there are numerous challenges, not least the language and financial issues, but there are recent positive examples, such as internships offered by INSEE, the French national statistical institute, to an EMOS student from other countries.

The commitment from the producer side to host internships and, like is the case in DESTATIS, the German statistical office, to mention EMOS as an advantage in vacancy notices, will be crucial for the attractiveness of EMOS in the future.

On the other hand, the producer side would like to see future EMOS graduates better equipped with skills and competences needed to manage a changing production system, including data science skills related to new and multiple data sources and modern tools. Here, universities will need to ensure that they can deliver by adapting their curricula to changing needs of future employers of EMOS graduates. By addressing the future skills needs, EMOS will be relevant and attractive for both employers and students.

Universities are also calling for common learning resources, including case studies, exercises with databases, simulations, games, quizzes, homework assignments, in particular for the learning outcomes where universities traditionally do not have sufficient teaching capacity unless they can rely on teaching staff from national statistical institutes (learning outcomes 1 and 5 mentioned above, as well as data science skills).

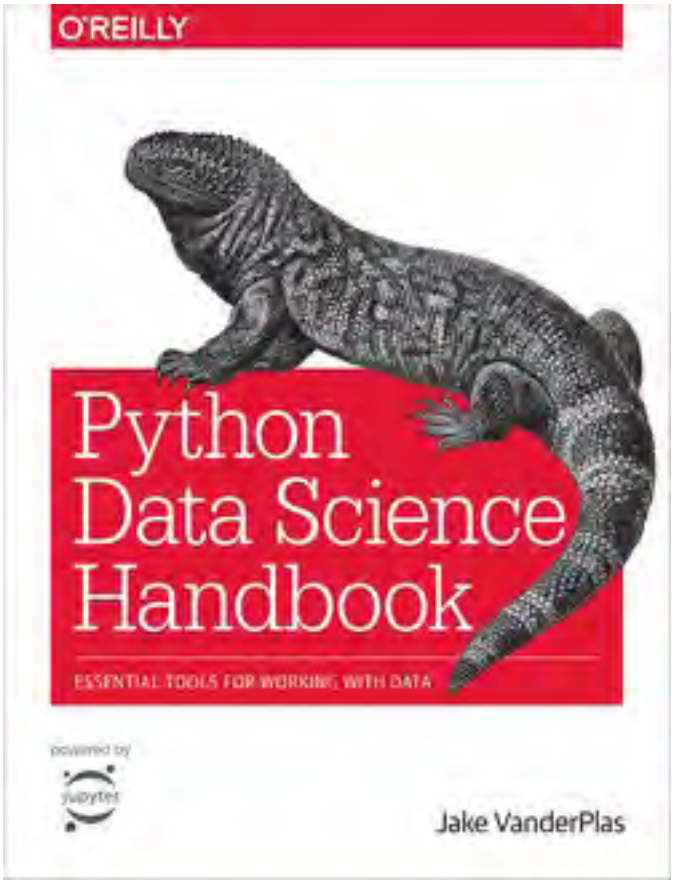
5. Conclusions

Since its launch, EMOS has grown from a simple idea to a network of 32 EMOS-labelled programmes in 19 countries across Europe. This in itself is a great achievement, which has only been possible thanks to the enthusiasm of all those involved. In fact, until now EMOS has been implemented with very little resources. And, as with all innovative projects, it takes more than enthusiasm for success, it takes serious commitment, hard work and resources as well. In the case of EMOS, the enthusiasm is definitely there, but the future level of ambition will be defined by the commitment of all stakeholders involved and the amount of resources they are ready to invest in EMOS. In the next year, a reflection on the level of ambition of EMOS will take place based on the progress made so far. The exciting journey started in 2009 has in fact just begun…

Python Data Science Handbook

by Jake VANDERPLAS (2016)

Alexis EIDELMAN[1]
Statisticien et data scientist public, administrateur hors classe de l'Insee



Livre (https://jakevdp.github.io/PythonDataScienceHandbook/)
Auteur : Jake VANDERPLAS
Édition : O'Reilly Media, Inc. – 2016
ISBN : 9781491912058

1. alexis.eidelman@travail.gouv.fr

DO YOU SPEAK PYTHON ?

La plupart du temps, le statisticien n'utilise qu'un seul langage de programmation. Même s'il connait parfois les rudiments de plusieurs d'entre eux, il s'est souvent spécialisé et n'en pratique qu'un seul. Pourtant, de la même manière que de plus en plus de personnes pratiquent plusieurs langues selon le contexte (familial, professionnel, voyages, etc.), le statisticien sera peut-être amené à utiliser plusieurs langages informatiques selon ce qu'il souhaite faire (collecte de données, traitement statistique, data visualisation, modélisation, etc.).

Un des langages intéressants aujourd'hui, à utiliser seul ou en complément d'autres, est le langage Python dont le livre *Python Data Science Handbook* constitue une référence pour initier le statisticien. Si la documentation des librairies et la communauté sur Python est conséquente et peut suffire, ce livre écrit par Jake VanderPlas, accessible librement[2], peut faciliter la prise en main et permettre de se « lancer » dans la découverte de Python.

Il est décomposé en quatre parties qui déroulent différents aspects du travail du statisticien/*datascientist*, plus une partie préliminaire consacrée à l'environnement Python utilisée dans le livre qui est essentielle. On pourra regretter toutefois que cette partie ne mentionne pas l'environnement Spyder qui est à mon sens l'environnement le plus pratique pour un usage « data » de Python.

Outre la partie sur les graphiques sur laquelle je ne m'étendrai pas, le livre présente les avantages de Python via trois entrées qui correspondent à trois niveaux de l'utilisation de Python et qui, sans peut-être le vouloir, correspondent à une chronologie du développement croissant de ce langage pour le traitement des données : calcul matriciel, manipulation de tableaux de données, *machine learning*.

Pour commencer, présentons en quelques mots le langage Python. Nommé en hommage aux Monty Python (et non pas en référence à un dangereux serpent), il se veut un langage généraliste mettant la lisibilité et la simplicité de sa syntaxe comme principe premier. Ce point, additionné au caractère *open source* du langage, lui a sans doute permis de devenir une référence pour le travail collaboratif. De ce fait, il a vu de nombreux *packages*/librairies se développer et est devenu grâce à eux un véritable couteau-suisse utilisé pour le traitement des données mais aussi pour les protocoles réseaux, pour des moteurs de recherches, pour les sites web et pour pratiquement tout ce que l'on peut faire avec un ordinateur. Le parti pris du *Python Data Science Handbook* est de ne pas s'attarder sur les bases du langage mais de renvoyer dans la préface à un ouvrage du même auteur qui permet d'apprendre pas à pas ses rudiments (fonctions, boucles, listes, etc.).

Langage de haut-niveau au sens informatique du terme, ce que Python gagne en simplicité/lisibilité il pourrait le perdre en performance par rapport à des langages comme C/C++, Fortran ou Perl. C'est sans compter sur différentes options permettant de retrouver cette puissance. A ce titre, un virage dans l'utilisation de Python pour le traitement de données a été le développement de Numpy, une librairie qui permet dans une syntaxe Python de faire du calcul matriciel avec la puissance de C. Cette librairie a permis à Python de se diffuser dans bon nombre de communautés scientifiques, en particulier dans l'univers de la physique. La partie de l'ouvrage qui présente Numpy décrit les bases de son utilisation et permet de comprendre comment accéder aux valeurs d'un tableau et la machinerie à l'œuvre lors des calculs, mais cette partie ne nécessite pas une lecture minutieuse car il n'est que très rarement nécessaire d'interagir directement avec cette librairie.

1. https://jakevdp.github.io/PythonDataScienceHandbook/

En effet, on se contente le plus souvent d'utiliser une autre librairie de Python, Pandas, qui utilise Numpy mais qui ajoute aux tableaux des noms de colonnes et prévoit des fonctions pour calculer des moyennes, pour regrouper les données selon une catégorie, etc. Cette librairie Pandas est « la » librairie de traitement de données statistiques. Le livre présente une introduction progressive assez proche de celle que l'on peut trouver dans la documentation de Pandas. On appréciera en particulier le passage sur les valeurs manquantes, les fusions de table ou encore sur les groupes. Cette partie comme le reste de l'ouvrage est très clair et mêle des textes d'explication et du code. Ce dernier peut d'ailleurs être exécuté depuis son ordinateur (le préliminaire l'explique), ce qui n'est pas sans vertu pédagogique.

Enfin, et c'est l'objet de la dernière partie, le langage Python est un langage de référence pour l'apprentissage automatique (*machine learning*). Développé par des communautés d'informaticiens, de statisticiens et d'experts en traitement d'image, l'apprentissage automatique s'est développé en Python et non dans des langages orientés vers les statistiques. Les librairies les plus utilisées du domaine sont écrites en Python et lorsque de nouveaux langages spécifiques se développent (Lua, Torch, TensorFlow, par exemple), ils sont très vite accompagnés d'une interface Python qui permet de garder la syntaxe et l'environnement Python tout en bénéficiant de la performance de ces langages dans leur domaine d'application. L'utilisation de Tensorflow est ainsi évoquée dans le livre peut-être trop rapidement pour les lecteurs qui souhaitent se pencher sur ce sujet. À sa décharge, l'ouvrage *Python Data Science Handbook* date de 2016, et s'il n'est pas obsolète et reste une excellente introduction, il ne couvre pas les progrès importants de ces dernières années dans le domaine du *machine learning*.

Le livre est un manuel de qualité mais s'adresse à des personnes qui ont déjà décidé d'apprendre le langage. Il ne répond pas – et ce n'est pas son rôle – à la question : quelles sont les qualités de Python qui pourraient conduire le statisticien à l'utiliser ? Une citation circule sur Python le présentant comme n'étant que le deuxième meilleur langage, mais dans tous les domaines. Je la trouve particulièrement pertinente. Certes, le statisticien spécialiste de son domaine préférera pour ses études utiliser le meilleur logiciel de statistique et il aura raison. Pourtant, dans bien des cas, le travail ne consiste pas seulement à ouvrir une base de données déjà nettoyée sur laquelle il n'y a plus que les calculs à réaliser et à présenter. Il est parfois – et même souvent – nécessaire de réaliser des opérations à la frontière et qui ne sont pas nécessairement statistiques. Un langage comme Python qui peut aisément gérer une chaine de traitement avec une gestion des messages d'erreur efficace, manipuler, convertir, mettre en ordre des données efficacement avec différents systèmes de stockage, nettoyer et gérer, par exemple, les chaines de caractères, réaliser des appels à des API ou à des sites web, faire de la représentation graphique, etc., devient alors un atout en particulier pour toute opération récurrente.

Python s'est développé loin du monde de la statistique et n'est devenu intéressant pour les statisticiens que récemment (l'indispensable librairie Pandas à moins de dix ans !). Puisque je tire le parallèle avec les langues parlées, je me risquerais à dire que Python est un peu l'anglais de la programmation. Parlé plus ou moins bien, il permet de faire le pont avec un écosystème très riche (informaticiens, physiciens, etc.) et de bénéficier de leurs apports via les librairies, mais aussi via les tutoriels ou forums d'échanges techniques. Ces éléments produits par des personnes du monde entier sont d'ailleurs pour beaucoup, comme le livre *Python Data Science Handbook*, en anglais.

Existe-t-il un avantage à commencer la séance de tirs au but au football ?



Luc ARRONDEL[1]
CNRS, PSE

Richard DUHAUTOIS[2]
CNAM-Lirsa et Ceet

Jean-François LASLIER[3]
CNRS, PSE

TITLE
The first-mover advantage in penalty shoot-outs: Really?

RÉSUMÉ
Dans cet article, nous analysons les séances de tirs au but lors des matchs de football dans trois compétitions : la Coupe de France, la Coupe de la Ligue et le Trophée des Champions. Nous nous intéressons aux effets psychologiques auxquels le joueur est soumis lors de ces séances : la « peur » de gagner, la « peur » de perdre et la « peur » de rattraper son adversaire. Notre principale conclusion est que la performance est affectée négativement à la fois par l'enjeu et par la difficulté de la situation mais nous ne trouvons aucun avantage à commencer (tirer en premier) la séance de tirs au but.

Mots-clés **:** *séances de tirs au but, avantage à tirer le premier, émotions, pression.*

ABSTRACT
The paper analyses sequences of penalty kicks during football shoot-outs in French cup competitions. We consider the psychological effects to which the kicker is subject: the "fear" of winning, the "fear" of losing, and the "fear" of catching up his opponent. Our main conclusion is that the performance (the probability of scoring) is negatively affected by both what is at stake and the difficulty of the situation. We find no advantage for the team that takes the first kick.

Keywords**:** *penalty shoot-outs, first-mover advantage, emotions, pressure.*

1. luc.arrondel@ens.fr
2. richard.duhautois@lecnam.net
3. jean-françois.laslier@ens.fr

1. Introduction

Les historiens datent la naissance du penalty à la toute fin du XIXe siècle. L'idée aurait été suggérée en 1890 par le gardien de but irlandais du *Milford Everton FC*, William McCrum, effrayé par la violence sévissant dans les surfaces de réparation (jusqu'à parfois causer la mort ![4]). Il imagine alors un coup de pied arrêté pour sanctionner l'équipe qui commet une faute. Après avoir imposé son idée au niveau local, il persuade la fédération irlandaise, dont il est membre, de la soumettre à l'*International Football Association Board* (*IFAB*), l'organe garant des règles du football depuis 1886. Malgré les réticences et les railleries des anglais de l'époque, le penalty fut introduit dans les Lois Officielles du Jeu en 1891, matérialisé par la quatorzième des dix-sept « lois du jeu » actuellement en vigueur. Pendant longtemps, le penalty pouvait être tiré de n'importe quel point le long de la ligne de « douze yards » et le gardien de but était autorisé à avancer jusqu'à « six yards » devant son but, les autres joueurs devant se positionner au moins « six yards » derrière le ballon. Le penalty « moderne » tel que nous le connaissons aujourd'hui a été instauré au début du XXe siècle.

Sa fonction première est de sanctionner une faute dans la surface de réparation adverse mais depuis 1970 pratiquement tous les matchs à élimination directe se terminent aux tirs au but (TAB) – une série de cinq penaltys en alternance – pour remplacer le tirage au sort lorsqu'un match se termine par un match nul. Pour les joueurs, le résultat de ces TAB dépend d'un nombre de facteurs, comme la compétence et la fatigue, mais aussi d'aspects psychologiques, telles que l'émotion et la pression (cf. infra et cf. Arrondel *et al.* (2019) pour une revue de littérature). Contrairement à ce que l'on entend souvent, une séance de TAB n'est pas une « loterie », à savoir une épreuve totalement aléatoire.

Ces séances de TAB ont par conséquent, comme les penaltys, fait l'objet de nombreuses analyses. De nombreux effets psychologiques ont été évoqués qui pourraient affecter le résultat final : le rang du tireur (être le premier ou le dernier tireur par exemple) ; la gestion de l'anxiété entre la fin des prolongations et les tirs au but ; être dans la dernière équipe à marquer avant la séance de tirs au but ; être ou ne pas être une star ; être d'une nationalité particulière ; avoir un maillot rouge ; etc. La question qui nous intéresse dans cet article est la suivante : a-t-on un avantage à commencer la séance de tirs au but ?

Quelques économistes se sont emparés de la question qui a déclenché une « controverse ». D'un côté, dans l'étude la plus connue publiée par Apesteguia et Palacios-Huertas (2010) avec un échantillon de 269 séries, l'équipe qui commence la séance remporte les TAB dans 60% des cas. De l'autre côté, Kocher *et al.* (2012), avec un échantillon plus grand, ne trouvent aucun avantage à commencer la séance. Les premiers justifient les résultats par la pression psychologique que ressentiraient les tireurs de la deuxième équipe. Étant donné la forte concurrence que subissent les footballeurs pour arriver au plus haut niveau, on peut se demander si cette pression psychologique que décrivent Apesteguia et Palacios-Huertas (2010) existe. Dans le championnat argentin, pendant la saison 1988-1989, la fédération a expérimenté une règle originale : chaque fois que le match se terminait par un match nul (environ 30% des rencontres), une séance de tirs au but avait lieu pour désigner le gagnant. Les séances de tirs au but étaient alors plus nombreuses et les équipes s'y entrainaient davantage. Les résultats montraient alors que l'équipe qui tirait en premier avait gagné dans 49,5% des cas. Autrement dit, aucune différence de réussite significative n'existe entre les équipes qui tiraient en premier et celles qui tiraient en second.

4. Un article de The Lancet datant du 22 avril 1899 souligne que 96 joueurs sont morts en jouant au football (et au rugby) pendant les huit années précédentes.

Nous avons donc décidé de rentrer dans cette controverse avec des données issues de compétitions en France. Nous avons analysé une série de séances de TAB tirées lors des matchs de Coupe de France, de Coupe de la Ligue et du Trophée des Champions, soit près de 250 séances pendant 15 ans. Nos résultats montrent que l'avantage à tirer le premier n'existe pas dans ces compétitions mais que d'autres aspects psychologiques apparaissent par contre plus importants.

2. Les penaltys et les tirs au but dans la littérature

L'impact de la pression psychologique sur la performance a été analysé dans de nombreux sports, de l'haltérophilie (Genakos et Pagliero, 2012) au golf (Hickman et Metz, 2015), en passant par le volley (Bozhinov et Grote, 2019) ou les échecs (Gonzalez-Diaz et Palacios-Huerta, 2016). Dans cette section, nous passons succinctement en revue quelques articles sur la stratégie des penaltys (pendant les matchs) et sur les séances de TAB[5].

2.1 La stratégie des penaltys

Commençons par une simple statistique : lors des matches des cinq grands championnats européens, le tireur convertit son penalty un peu plus de trois fois sur quatre en moyenne et ce pourcentage est particulièrement proche d'un pays à l'autre (tableau 1). Mais lors des entrainements, les joueurs réussissent 90% des penaltys suggérant que des aspects psychologiques sont à l'œuvre en compétition en défaveur du tireur (Chiappori *et al.*, 2002).

Tableau 1 – *Penaltys dans les cinq grands championnats européens entre les saisons 2006/2007 et 2015/2016*

Championnat	Nombre de penaltys	Nombre de penaltys marqués	Part des penaltys marqués (%)
Premier League (Angleterre)	948	739	77,95
Bundesliga (Allemagne)	745	571	76,64
La Liga (Espagne)	1089	831	76,31
Serie A (Italie)	1239	945	76,27
Ligue 1 (France)	931	708	76,05

Sources : Ligues Nationales

La motivation première de l'article de Chiappori *et al.* (2002) est d'étudier le comportement stratégique des tireurs de penalty et des gardiens de but en se référant à une sous-discipline de l'économie, la théorie des jeux. Leur modèle est un problème typique de théorie des jeux non coopératif où un tireur se trouve face à un gardien. Les questions sont alors les suivantes : pour le joueur, où dois-je tirer sachant que le gardien choisit un côté et pour le « goal », où dois-je plonger sachant que le tireur choisit un côté. Le jeu est à somme nulle puisque, soit le tireur marque le penalty, soit il le rate (arrêt du gardien ou tir à côté). La stratégie de chacun des protagonistes dépend donc de ce qu'il pense que l'autre va faire et on peut montrer que la décision optimale est de tirer de façon purement aléatoire en tenant compte néanmoins de certains facteurs physiologiques (droitier ou gaucher). Les auteurs définissent ainsi le « côté naturel » d'un tireur, c'est-à-dire le côté où c'est le plus simple de tirer : à gauche pour un joueur droitier et à droite pour un gaucher[6]. Pour maximiser la probabilité de marquer son penalty (on parle de stratégie mixte), le tireur devra tirer 60% de ses tirs de son côté naturel (celui où il est le plus à l'aise) et 40% du côté opposé.

5. Voir Arrondel et al. (2019) pour plus de détails.
6. Autrement dit, les joueurs croisent leurs tirs.

Pour tester les prédictions de leur modèle, les auteurs observent 459 penaltys tirés en *Ligue 1* française et en *Serie A* italienne à la fin des années 1990. Leurs données montrent que 45% des penaltys sont tirés du côté naturel, 17% au milieu et 38% du côté non naturel. Les gardiens plongent du côté naturel du joueur dans 57% des cas et du côté opposé dans 41% des cas ; ils ne sont restés au centre de leur but que dans 2% des cas. Ces résultats empiriques correspondent en partie aux prédictions du modèle théorique : par exemple, les tireurs tirent au milieu plus souvent que les gardiens restent au centre du but et les gardiens plongent plus souvent sur le côté naturel du joueur que sur le côté non naturel. Palacios-Huerta (2003) estime les mêmes probabilités en utilisant un échantillon plus grand de 1 417 penaltys observés principalement dans les championnats d'Espagne (*La Liga*), d'Angleterre (*English Premier League*) et d'Italie (*Serie A*). Ses résultats confirment dans les grandes lignes ceux de l'étude précédente : les joueurs tirent plus souvent de leur côté naturel (53% contre 39% du côté non naturel et 8% au milieu). Les gardiens de but plongent du côté naturel du joueur dans 58% des cas et du côté opposé dans 40% des cas. La probabilité de rester au centre du but est la même : environ 2%.

Le gardien reste donc très rarement au milieu de son but pour essayer d'arrêter un penalty. Bar-Eli et ses co-auteurs (2007, 2009, 2011) s'intéressent à ce comportement dans le même cadre de théorie des jeux que les auteurs précédents. Ils émettent l'hypothèse que le gardien préfère plonger plutôt que de rester au centre de son but en raison d'un « biais comportemental » : les supporters n'aiment pas voir les gardiens ne pas plonger et ces derniers le savent. Pour conforter leur hypothèse, les auteurs ont visionné un échantillon de penaltys et ont interrogé des gardiens de but professionnels pour obtenir des informations complémentaires sur leurs comportements. Leurs conclusions montrent que : (1) les gardiens plongent sur les côtés plus souvent qu'ils ne le devraient ; (2) ce comportement non optimal est perçu comme la « norme » par les gardiens de but ; (3) sans surprise, les tirs dans la partie supérieure du but sont les plus difficiles à arrêter ; (4) les gardiens de but sont plus satisfaits quand ils arrêtent un tir en hauteur. Comme les études précédentes, ils montrent que leurs données ne contredisent pas les prédictions d'un équilibre de Nash en stratégie mixte. Jantschgia and Nax (2020) montrent que cet équilibre ne serait par contre pas observé lors des séances de TAB : dans ce cas, le choix d'un tireur dépendrait de celui des tireurs précédents.

2.2 Les analyses des séances de tirs au but

Les tirs au but sont une succession de penaltys en alternance pour départager les équipes qui n'ont pas réussi à le faire pendant le temps réglementaire et (souvent) les prolongations. De nombreuses études ont mis certains faits en évidence lors de ces séances de TAB :

1. Les tirs au but sont moins réussis en fin de séance qu'en début de séance (Jordet et al., 2006). Séparer pression plus forte ou moindre compétence des tireurs n'est cependant pas facile. Une certaine compétence est nécessaire car les auteurs constatent que dans l'exercice des TAB, ce sont les attaquants qui marquent en proportion le plus, puis les milieux de terrain et enfin les défenseurs.
2. L'équipe qui a marqué la dernière avant la séance – et qui donc est revenue au score dans le match – a une probabilité plus élevée de gagner la séance de TAB (61%, Littleton, 2016).
3. La gestion de l'anxiété entre la fin du match et le début des TAB est un facteur important de réussite (Jordet et Elferink-Gemser, 2012).
4. Les meilleurs joueurs sont plus susceptibles de rater leur tir au but (Jordet, 2009) : 65% de réussite contre près de 74% pour les autres joueurs. Auraient-ils plus de pression en raison de leur statut particulier ? McGarry et Franks (2000) suggèrent que les meilleurs joueurs devraient tirer en cinquième position et plus généralement en ordre inverse de leur compétence à tirer.
5. Les effets des caractéristiques des « cultures nationales » (Billsberry et al. (2007) : ce sont

6. pourtant incontestablement intéressé par le caractère fructueux de l'approche bayésienne dans sa propre recherche puisqu'il a été le directeur de thèse de Christian Robert.

les nations dites plus « individualistes » qui auraient le plus de difficultés à gagner (plus de responsabilité individuelle ?). C'est par cette dimension culturelle que l'on pourrait expliquer la « malchance » des anglais dans cet exercice (Jordet, 2009).

6. La couleur du maillot serait aussi une autre dimension psychologique à prendre en compte : Greenlees *et al.* (2008) affirme que porter un maillot rouge impressionne davantage les gardiens.

Une des règles de base de ces séances de tirs au but est que les équipes tirent alternativement (AB puis AB...) après qu'un tirage au sort (depuis 2003) a donné la possibilité au capitaine de l'équipe ayant gagné le tirage de choisir de tirer en premier ou en second. Le problème est que l'équipe tirant la première aurait un avantage psychologique qui lui permettrait de gagner plus souvent. En d'autres termes, le second tireur serait en permanence sous « pression », ce qui lui ferait rater son tir au but plus que la moyenne. C'est ce qu'énoncent les travaux de Apesteguia et Palacios-Huerta (2010). Selon ces auteurs, les statistiques révèlent un avantage énorme à l'équipe du premier tireur, de l'ordre de 60/40 : en d'autres termes, l'équipe débutant la séance de TAB aurait six chances sur dix de gagner l'épreuve, la seconde seulement quatre sur dix. Palacios-Huerta a d'ailleurs été reçu par l'*International Board*, le garant des lois du jeu, pour envisager une réforme des séances de TAB en changeant l'ordre des tireurs. Il semble qu'il ait été entendu puisqu'une nouvelle règle, à l'instar du *tie-break* au tennis, a été testée à l'occasion de l'Euro U17 féminin 2017 en République Tchèque : un tireur de l'équipe A se présente, puis deux tireurs de l'équipe B suivis d'un nouveau tireur de l'équipe A. Ce nouvel ordre est censé rétablir l'équilibre compétitif entre les deux équipes.

Le problème de la conclusion d'Apesteguia et Palacios-Huerta justifiant cette réforme est qu'une étude postérieure de Kocher *et al.* (2012) a montré, avec un échantillon deux fois plus grand (540 séances contre 269), que l'écart entre la probabilité de gagner de la première équipe et celle de la seconde était inférieur (53/47) et surtout statistiquement non significatif. Cependant, Palacios-Huertas (2014) augmente à son tour l'échantillon à 1 001 tirs au but et retrouve son premier résultat.

Vandebroek *et al.* (2018) cherchent alors à comprendre pourquoi les études empiriques sont en désaccord et essaient de réconcilier les résultats. Ils soutiennent l'hypothèse de Palacios-Huertas que la pression psychologique augmente la probabilité que la première équipe gagne la séance de TAB et montrent que l'échantillon utilisé par Kocher et al. (2012) est de taille trop faible pour trouver un effet significatif lorsque la probabilité de gagner est de 53%.

Il est surprenant que cette controverse basée sur une simple statistique ne soit pas tranchée puisqu'il s'agit d'une fréquence empirique observable : faiblesse de l'échantillon, biais de sélection ? Il est donc important de poursuivre l'analyse : c'est ce que nous avons fait personnellement sur des données françaises.

3. Les séances de TAB en France : les Coupes

Les données proviennent de trois compétitions françaises : essentiellement la Coupe de France et la Coupe de la Ligue, et quelques matchs du Trophée des Champions (vainqueur du championnat contre le vainqueur de Coupe de France). La Coupe de France est une compétition nationale ouverte à tous les clubs amateurs et professionnels et la Coupe de la Ligue n'est ouverte qu'aux clubs professionnels (en Ligue 1 et Ligue 2, ainsi que certains clubs de National 1, la troisième division). Les données se composent de 252 séances de TAB (135 pour la Coupe de France entre 2007 et 2017, 110 pour la Coupe de la Ligue entre 2001 et 2018 et les sept séances du Trophée des Champions depuis sa création en 1995). Ces données ont été collectées sur différents sites, dont celui de la LFP (Ligue de football professionnel) et le journal *L'Équipe*. Pour chaque séance, nous avons recueilli des informations sur la date de la séance, le nom des

équipes tirant en premier et en second, le résultat final des TAB, le nombre de personnes dans le stade et l'emplacement géographique du match[7]. Le tableau 2 présente ces statistiques.

Tableau 2 – *Description de l'échantillon*

Compétition	Saison	Nombre de séances
Coupe de France	*2007-2017*	*135*
64e	2007-2018	46
32e	2007-2019	52
16e	2007-2020	21
8e	2007-2021	8
quarts	2007-2022	8
Coupe de la Ligue	*2001-2018*	*110*
Tours préliminaires	2001-2018	53
16e	2001-2019	26
8e	2001-2020	16
quarts	2001-2021	10
demies	2001-2022	4
finale	2001-2023	1
Trophée des Champions	*1995-2017*	*7*

Sources principales : LFP et Léquipe.fr.

La figure 1 présente les probabilités de gagner lorsque l'équipe commence la séance de TAB. En globalité, la répartition est plus ou moins équiprobable : la probabilité de victoire est de 50,4% pour l'équipe qui commence la séance et de 49,6% pour l'équipe qui tire en second. Les résultats ne changent pas par compétition : les chiffres sont de 51,1% contre 48,9% pour la Coupe de France et 49,1% contre 50,9% pour la Coupe de la Ligue[8]. Aucun de ces ratios n'est statistiquement différent de 50% et une analyse toutes choses égales par ailleurs montre que la seule variable qui joue un rôle dans la probabilité de victoire est le fait de jouer à un niveau plus élevé (Arrondel *et al.*, 2019). Pour résumer, nos résultats descriptifs montrent qu'il n'y a pas d'avantage, au moins pour les compétitions en France, à tirer le premier lors d'une séance de tirs au but.

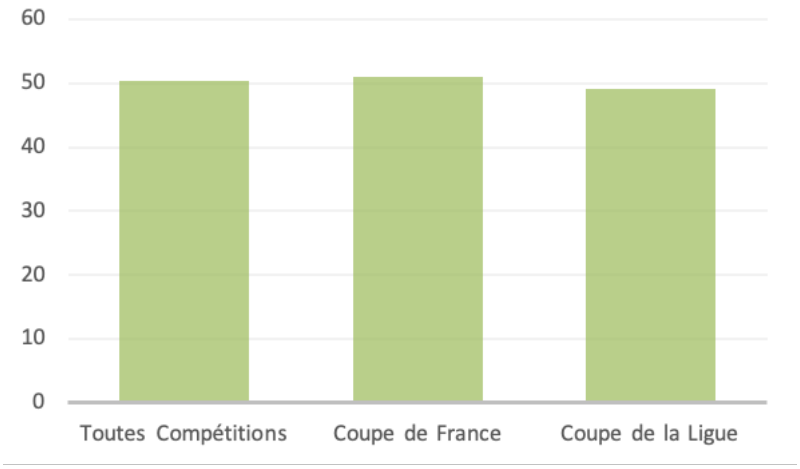


Figure 1 – *Fréquences (en %) des victoires lorsque l'équipe tire la première*

7. Treize séances de TAB sur les 252 ne contiennent pas l'issue de chaque tir.
8. Les sept observations concernant le Trophée des Champions ne nous permettent pas de dresser un bilan.

En revanche, nos données confirment en grande partie deux conclusions communes à la littérature (figure 2)[9] :

1. Le taux de réussite en moyenne est plus faible lors des tirs au but qu'en cours de match : 73,1% contre plus de 76%.
2. La probabilité de marquer le penalty décroit avec l'avancée de la séance : près de 80% sont marqués pour le premier tir contre moins de 70% pour le cinquième tir.

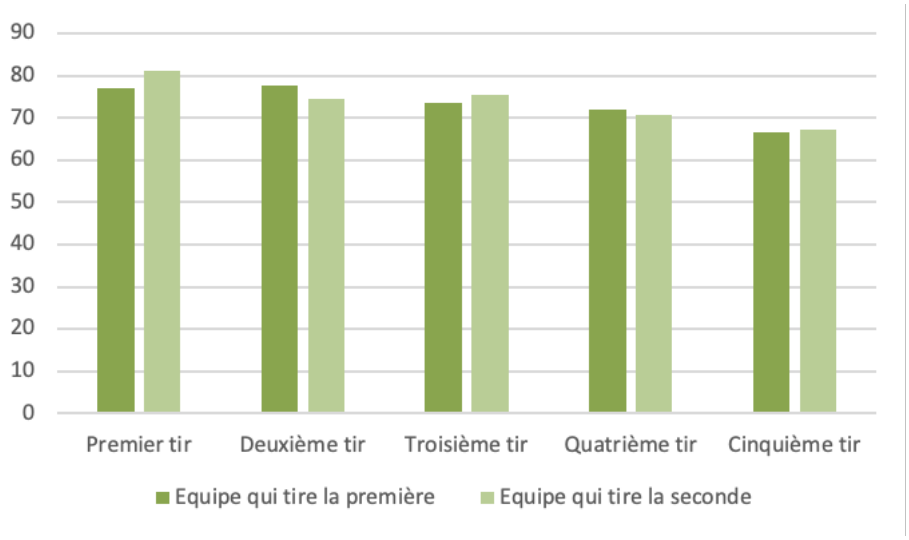


Figure 2 – *Fréquences (en %) des réussites lors des différents tirs*

Ces deux résultats s'expliquent tout simplement par le fait que tous les joueurs ne sont pas des spécialistes des penaltys et que les équipes mettent généralement leurs meilleurs tireurs au début de la séance pour éviter d'être distancées (McGarry et Franks, 2000).

4. « L'enjeu » et la « pression » au moment de tirer

Puisque nos données ne semblent pas soutenir l'hypothèse d'un avantage à tirer le premier, nous testons une hypothèse concurrente. D'une part, nous considérons que la pression psychologique n'entre en jeu qu'après le premier tour, lorsque le nombre de tirs est identique. Le premier tireur de la deuxième équipe ne ressentirait pas d'émotion particulière à tirer après le premier tireur de l'autre équipe, même si ce dernier a marqué. D'autre part, nous considérons que si une équipe est en retard lorsque le nombre de tirs est identique pour les deux équipes, le prochain tireur peut alors ressentir une pression car manquer son tir augmenterait le risque de perdre pour son équipe.

Plus formellement, Palacios-Huertas et ses coauteurs testent simplement le fait d'avoir un but de retard sur la probabilité de marquer ($Y = 0$ ou $Y = 1$) avec ou sans variables de contrôle (X). Ainsi, la probabilité de marquer s'écrit

$$\mathrm{P}(Y = 1) = F(p_0 + \alpha \Delta\text{but} + \beta X),$$

avec $\mathrm{F}(\cdot)$ une fonction de répartition quelconque, p_0 une constante, et α et β des paramètres à estimer. Dans notre cas, nous estimons l'effet d'avoir un but de retard à nombre de tirs équivalents, c'est-à-dire

$$\mathrm{P(Y=1)=F}(p_0' + \alpha' \, [\Delta\text{but}(+1 \text{ ou } 0)] + \beta X).$$

On rajoute +1 à l'équipe qui tire en second et 0 à l'équipe qui tire en premier. X représente le

9. Puisqu'il nous manque l'information sur la réussite de quelques penaltys, nous ne pouvons reconstituer entièrement que 239 des 252 séances de TAB de l'échantillon.

niveau des équipes, c'est-à-dire la division dans laquelle les équipes jouent (Ligue 1, Ligue 2, National, etc.).

Le tableau 3 présente les résultats des estimations de la première équation (deux premières colonnes) et de la seconde (deux dernières colonnes) avec deux modèles à effets fixes, linéaire et Logit. Les résultats montrent que la simple différence de buts (comme introduit par Palacios-Huerta) n'a aucun effet sur la probabilité de marquer mais qu'en revanche la différence de buts à nombre de tirs équivalents a bien un effet positif. On retrouve également, à l'instar des statistiques descriptives, l'effet négatif de l'ordre des tireurs.

Tableau 3 – *Probabilité de marquer un penalty lors d'une séance de TAB*

	Effets fixes (Linéaire)	Effets fixes (Logit)	Effets fixes (Linéaire)	Effets fixes (Logit)
Différence de buts	-0,001 (0,013)	-0,041 (0,067)		
Différence de buts (Nombre de tirs équivalents)			0,083*** (0,013)	0,370*** (0,065)
Rang du pénalty	-0,031*** (0,005)	-0,163*** (0,028)	-0,032*** (0,005)	-0,170*** (0,029)
Nombre de Tirs Nombre de Séries	2 504 239	2 504 239	2 504 239	2 504 239

En utilisant une spécification un peu plus sophistiquée de la seconde équation, on considère trois situations. La première correspond au « tir de rattrapage » : lorsque le joueur tire avec un but de retard et l'équipe a un penalty de moins. La seconde correspond à la situation où, à nombre de tirs équivalents, l'équipe du tireur a un but d'avance (tir de « break »). Enfin, la troisième situation correspond à celle où, à nombre de tirs équivalents, l'équipe du tireur a un but de moins (tir de « survie »).

La figure 3 présente les probabilités de marquer dans les trois situations. Nous pouvons voir qu'il y a un désavantage psychologique à rattraper l'adversaire et à tirer pour la « survie » de l'équipe car les taux de réussite se situent respectivement à 70,2% et à 70,8%. En revanche, avoir un but d'avance est plutôt une situation relativement confortable – rien n'est perdu même si le joueur rate son tir – puisque le taux de réussite est plus élevé (83,6%).

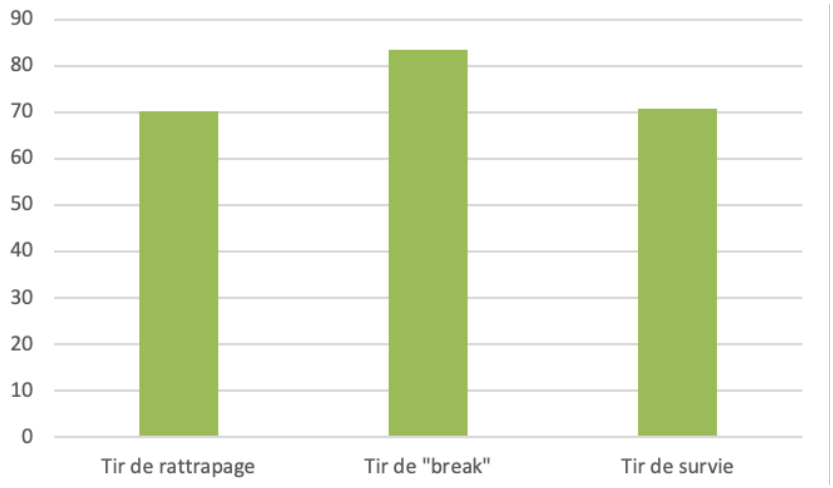


Figure 3 – *Fréquences (en %) des réussites en fonction de la nature des tirs au but*

Tableau 4 – *Probabilité de marquer un penalty lors d'une séance de TAB*

	Effets fixes (Linéaire)	Effets fixes (Logit)
Survie	-0,112*** (0,030)	-0,663*** (0,170)
Rattrapage	-0,249*** (0,024)	-1,340*** (0,146)
Rang du pénalty	-0,041*** (0,005)	-0,220*** (0,030)
Nombre de Tirs Nombre de Séries	2 504 239	2 504 239

Note : La survie et le rattrapage s'expriment par rapport au tir de « Break ». Le niveau des équipes est pris en compte mais non répertorié.

Ces effets bruts sont corroborés par les estimations économétriques en tenant compte de l'hétérogénéité des séances, du rang du penalty et du niveau des équipes (tableau 4). Un parallèle pourrait être établi avec des études concernant l'influence du « stress » sur la prise de risque : les individus prennent plus de risques lorsqu'ils ont « tout » à perdre mais prennent moins de risques lorsqu'ils ont « tout » à gagner (Porcelli et Delgado, 2009). « L'angoisse du joueur au moment du penalty » est donc beaucoup plus complexe que la simple hypothèse d'être dans l'équipe qui commence la séance de TAB.

5. Conclusion

Les séances de tirs au but à l'issue d'un match de football permettent à de nombreux chercheurs d'analyser le comportement et la stratégie des joueurs. Contrairement aux hypothèses des modèles standards de la théorie économique, le tireur de penalty lors d'une séance de TAB n'arbitre pas entre un coût et un bénéfice (entre l'effort fourni et les chances de gain) en fonction de ses préférences (aversion pour le risque notamment) : puisque la séance se déroule à la fin du match, les joueurs ont « tout à gagner » ou « tout à perdre », ils font simplement leur maximum et ce sont les facteurs physiques et psychologiques qui sont à l'œuvre. Ainsi, dans cet article, nous avons notamment montré que l'enjeu et le risque de perdre réduisent les probabilités de réussite : même les tireurs expérimentés sont moins susceptibles de marquer lors de ces exercices quand l'enjeu ou le risque de perdre sont forts. Jordet (2009) montre que les « superstars », souvent tireurs de penaltys en match, ont moins de chance de réussir que les autres joueurs lors des séances de tirs au but, du fait de la pression liée à leur statut. Au contraire, lorsque les joueurs sont dans une situation favorable, la probabilité de réussite est plus élevée.

Une illustration de ces divers effets psychologiques est bien résumée par Luis Fernandez qui commente son TAB donnant la victoire à l'équipe de France contre le Brésil en quart de finale de la Coupe du monde de 1986 : « Je me suis toujours mis en cinquième position quand je devais faire les tirs au but. Déjà avec le PSG. Quand Henri Michel a demandé, j'ai dit que je me mettais en cinquième position. Michel (Platini) venait de rater le sien, Julio Cesar aussi. Je savais que j'avais la qualification dans les pieds. J'y suis allé tranquillement, sans me précipiter. J'ai attendu que le gardien aille dans les buts. J'avais l'impression qu'il voulait me déstabiliser. Il faut ensuite faire le vide et savoir où on veut le tirer. J'étais relativement décontracté et détendu. J'étais sûr de mon fait. Ce n'était pas dans ma nature de paniquer. J'étais fort mentalement. Quand je me suis lancé, j'avais un endroit. Je me suis dit : Tu tires là ». L'enjeu était fort mais la pression plus

faible (s'il ratait la France restait en jeu).

Finalement, à propos de la question de savoir si l'équipe qui tire en premier est avantagée, il faut noter que les facteurs étudiés jouent dans des sens opposés. La difficulté empirique est due à la faiblesse – réelle – des échantillons, qui rend possible les biais de sélection et ouvre la voie aux biais de publication et aux controverses. Les « big data » vont sans doute trancher la question dans un avenir proche. Comme toujours, « les données finiront par parler ». La société d'analyse des performances sportives, *InStat*, a ainsi récemment publié un rapport (Instat, 2019) dans lequel elle analyse les séquences de 2 000 séries de TAB : résultat, 51,5% de chance de gagner pour l'équipe qui tire en premier.

Références

Arrondel L., R. Duhautois, and J.-F. Laslier (2019), « Decision Under Psychological Pressure: The Shooter's Anxiety at the Penalty Kick », *Journal of Economic Psychology*, vol. 70, pp. 22-35.

Apesteguia J. and I. Palacios-Huerta (2010), « Psychological pressure in competitive environments: evidence from a randomized natural experiment », *American Economic Review*, vol. 100, n° 5, pp. 2548–2564.

Azar O. H. and M. Bar-Eli (2011), « Do soccer players play the mixed-strategy Nash equilibrium? », *Applied Economics*, vol. 43, n° 25, pp. 3591-3601.

Bar-Eli M. and O. H. Azar (2009), « Penalty kicks in soccer: an empirical analysis of shooting strategies and goalkeepers' preferences », *Soccer and Society*, vol. 10, n° 2, pp. 183-191.

Bar-Eli M., O. H. Azar, I. Ritov, Y. Keidar-Levin, and G. Schein (2007), « Action bias among elite soccer goalkeepers: The case of penalty kicks », *Journal of economic psychology*, vol. 28, n° 5, pp. 606-621.

Billsberry J., P. Nelson, N. Van Meurs, and G. Edwards (2007), « Are penalty shoot-outs racist? », *Journal of Sports Science and Medicine*, vol. 6, n° 10, p. 98.

Bozhinov V. and N. Grote (2019), « Performance under Pressure on the Court: Evidence from Professional Volleyball », *Working Papers 1901*, Gutenberg School of Management and Economics, Johannes Gutenberg-Universität Mainz.

Chiappori P. A., S. Levitt, and T. Groseclose (2002), « Testing mixed-strategy equilibria when players are heterogeneous: The case of penalty kicks in soccer », *American Economic Review*, vol. 92, n° 4, pp. 1138-1151.

Genakos C. and M. Pagliero (2012), « Interim rank, risk taking, and performance in dynamic tournaments », *Journal of Political Economy*, vol. 120, n° 4, pp. 782-813.

Gonzalez-Diaz J. and I. Palacios-Huerta. (2016), « Cognitive performance in competitive environments: Evidence from a natural experiment », *Journal of Publics Economics*, vol. 139, pp. 40-52.

Greenlees I., A. Leyland, R. Thelwell, and W. Filby (2008), « Soccer penalty takers' uniform colour and pre-penalty kick gaze affect the impressions formed of them by opposing goalkeepers », *Journal of Sports Sciences*, vol. 26, n° 6, pp. 569-576.

Hickman D. C. and N. E. Metz (2015), « The impact of pressure on performance: Evidence from the PGA TOUR », *Journal of Economic Behavior & Organization*, vol. 116, pp. 319-330.

Instat (2019), *Penalties: Ultimate Guidelines*.

Jantschgia S. and H. Nax (2020), « Minimax on football penalties? Not at shoot out! A comment », mimeo.

Jordet G., E. Hartman, C. Visscher, and K. A. Lemmink (2007), « Kicks from the penalty mark in soccer: The roles of stress, skill, and fatigue for kick outcomes », *Journal of Sports Sciences*, vol. 25, n° 2, pp. 121-129.

Jordet G. and M. T. Elferink-Gemser (2012), « Stress, coping, and emotions on the world stage: The experience of participating in a major soccer tournament penalty shootout », *Journal of Applied Sport Psychology*, vol. 24, n° 1, pp. 73-91.

Jordet G., M. T. Elferink-Gemser, K. A. Lemmink, and Visscher (2006), « The "Russian roulette" of soccer?: Perceived control and anxiety in a major tournament penalty shootout », *International Journal of Sport Psychology*, vol. 37, n° 2/3, pp. 281-298.

Jordet G. (2009), « Why do English players fail in soccer penalty shootouts? A study of team status, self-regulation, and choking under pressure », *Journal of sports sciences*, vol. 27, n° 2, pp. 97-106.

Kocher M., M. V. Lenz, and M. Sutter (2012), « Psychological pressure in competitive environments: new evidence from randomized natural experiments », *Management Science*, vol. 58, n° 8, pp. 1585-1591.

McGarry T. and I. M. Franks (2000), « On winning the penalty shoot-out in soccer », *Journal of Sports Sciences*, vol. 18, n° 6, pp. 401-409.

Palacios-Huerta I. (2014), *Beautiful Game Theory*, Princeton University Press.

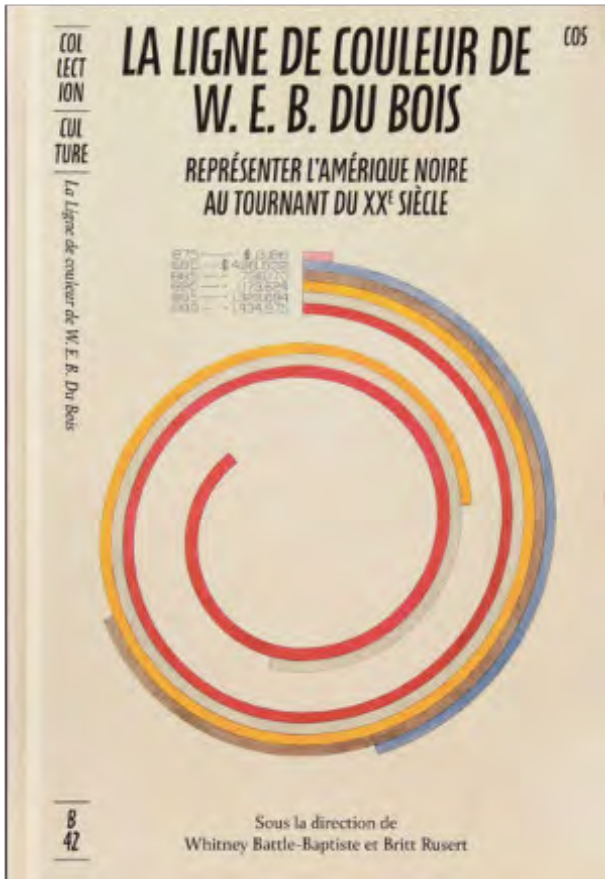
Palacios-Huerta I. (2003), « Professionals play minimax », *The Review of Economic Studies*, vol. 70, n° 2, pp. 395-415.

Vandebroek T. P., B. T. McCann, and G. Vroom (2018), « Modeling the Effects of Psychological Pressure on First-Mover Advantage in Competitive Interactions: The Case of Penalty Shoot-Outs », *Journal of Sports Economics*, vol. 19, n° 5, pp. 725-754.

La ligne de couleur de W. E. B. Du Bois
Représenter l'Amérique noire au tournant du XXe siècle

Sous la direction de Whitney BATTLE-BAPTISTE et Britt RUSERT (2019)

Antoine ROLLAND[1]
Université Lumière Lyon 2



Livre (144 pages – Traduit de l'anglais par Julia BURTIN ZORTEA)
Direction : Whitney BATTLE-BAPTISTE et Britt RUSERT
Édition : Éditions B42 (Collection : Culture) – 2019 (2018 pour l'édition américaine)
ISBN : 9782490077229

1. antoine.rolland@univ-lyon2.fr

Ce petit livre (144 pages) est remarquable à plus d'un titre. Tout d'abord, il est remarquable par la personne à qui il est consacré. Il met à l'honneur le sociologue William Edward Burghardt Du Bois, dit W. E. B. Du Bois. Né en 1863 dans le Massachusetts (États-Unis), W. E. B. Du Bois est un universitaire, sociologue, infatigable militant de la cause afro-américaine au tournant du XX^{e} siècle. Il est considéré aujourd'hui comme un des « pères oubliés de la sociologie moderne »[2]. Figure de proue de l'école de sociologie de l'université d'Atlanta, il s'attache particulièrement à l'émancipation des afro-américains, quelques décennies après la fin de la guerre de sécession et l'abolition de l'esclavage.

Ce livre est remarquable ensuite par le focus mis sur un travail particulier de W. E. B. Du Bois et son équipe. À l'occasion de l'exposition universelle de Paris en 1900, ceux-ci ont proposé une présentation graphique de statistiques relatives à la situation des afro-américains dans l'état de Géorgie, et plus généralement aux États-Unis. Cette présentation était hébergée dans le pavillon américain. Les représentations graphiques s'appuient sur des statistiques simples et efficaces, convenablement choisies par W. E. B. Du Bois pour servir son propos. La population afro-américaine apparait comme étant très diversifiée, de plus en plus éduquée, mais toujours pauvre et discriminée dans l'Amérique de 1900 malgré la fin de l'esclavage. Ces représentations graphiques ont été effectuées directement à la main sur de grandes feuilles de papier (56x71cm), au crayon, à l'encre et à la gouache, à partir de formes géométriques simples. Cette exposition est d'une modernité et d'une inventivité remarquable en terme de data-visualisation. Les contraintes de moyen et de place, alliées aux choix du discours amène à une justesse de représentation telle que la « ligne de partage des couleurs » au sein de l'Amérique apparait clairement. Elle est particulièrement soulignée par le choix des couleurs (noir, brun, rouge…) qui ajoute, sur certains graphes, une charge émotionnelle certaine.

Enfin ce volume est remarquable en ce sens que c'est un « beau livre », malgré sa taille réduite (16x24cm) mais adaptée au format initial des planches. Il reproduit fidèlement l'intégralité des planches et graphiques de l'exposition. Les couleurs sont très bien rendues, la traduction française est agréable à lire. C'est un livre de très grande qualité.

Le livre est divisé en trois parties. Une première partie est composée de quatre chapitres de présentation historique des travaux de W. E. B. Du Bois, et de cette exposition en particulier. Une deuxième partie présente les 36 planches de l'étude sociale sur « les Nègres de Géorgie ». Cartes chloroplèthes, diagrammes en bâtons verticaux, horizontaux, en valeurs absolues ou empilés à 100%, courbes chronologiques, diagrammes en secteurs, et bien sûr les « lignes sinueuses », spécialités de W. E. B. Du Bois : si la barre d'un diagramme en bâtons est trop longue pour tenir sur la page, W. E. B. Du Bois n'hésite pas à la replier ou l'enrouler en spirale, comme sur l'exemple donné sur la couverture. Ce graphique spécifique représente de manière très expressive l'incroyable augmentation de la valeur des biens mobiliers des Noirs de Géorgie entre 1875 et 1899. Les statistiques proposées sont des statistiques socio-démographiques et économiques : populations totales et par comté, pyramide des âges, illettrisme, nombre d'enseignants, patrimoine, métiers effectués, budgets familiaux…

Enfin la troisième partie est plus générale, et s'attache à décrire la population afro-américaine sur l'entièreté du territoire des États-Unis à travers 26 autres planches, sur des thèmes et avec des représentations semblables à l'exposition spécifique sur la Géorgie.

En résumé, ce petit livre est indispensable dans la bibliothèque de tout sociologue, statisticien, data-visualiseur ou simple esprit curieux. Les deux graphes ci-dessous, ainsi que leurs commentaires directement extraits du livre, sauront certainement vous donner envie d'en savoir plus !

2. D'après le quatrième de couverture du livre.

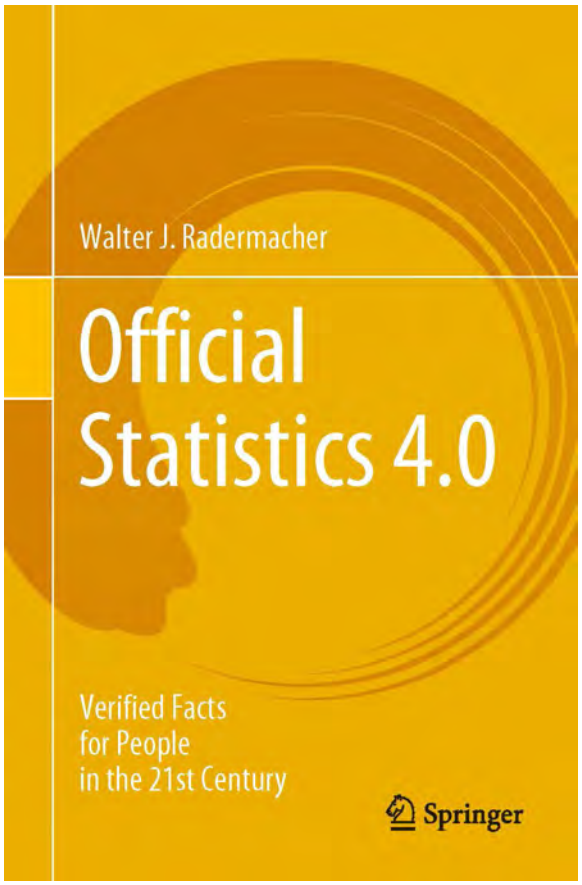
« Planche 10 – Condition conjugale :
Ce diagramme tricolore représente trois paires de bâtons empilés symbolisant les différents statuts conjugaux des Allemands et des Africains-Américains : célibataires en rouge, mariés en jaune, et veufs ou divorcés en vert. Ces bâtons sont également divisés par tranche d'âge. Du Bois a choisi l'Allemagne voisine lors de l'Exposition Universelle de Paris comme point de comparaison afin de valoriser les Africains-Américains auprès d'un public étranger mais aussi étatsunien. C'est à cette fin qu'il construit une relation graphique entre la population africaine-américaine et la population, principalement blanche, de l'une des premières puissances européennes. »

« Planche 12 – Esclaves et Nègres affranchis :
Ce graphique en aires, qui se lit de haut en bas, soit chronologiquement, est bordé sur l'un des côtés par une ligne nette, et de l'autre par une ligne dentelée. Sur la gauche, une profonde couleur noire représente les esclaves en Géorgie de 1790 à 1870. Sur la droite, la figure géométrique qui se dessine représente la hausse, le déclin puis la hausse à nouveau du pourcentage de Noirs affranchis. Un titre sobrement composé domine ce diagramme compact, renforçant l'impact visuel de l'image. »

Official Statistics 4.0
Verified Facts for People in the 21st Century

de
Walter J. RADERMACHER
(2020)

Thomas AMOSSÉ[1]
Cnam, Lise, CEET



Livre (158 pages)
Auteur : Walter J. RADERMACHER
Édition : Springer International Publishing– 2020
ISBN : 978-3-030-31491-0

1. thomas.amosse@lecnam.net

Walter J. Radermacher est un statisticien et économiste allemand qui a occupé dans son domaine d'éminentes fonctions au niveau national et européen, ayant présidé *Destatis* (*Deutschland Statistisches Bundesamt*) avant de prendre la direction générale d'*Eurostat* de 2008 à 2016. Fort de cette expérience et de la connaissance d'une vaste littérature, il présente dans cet ouvrage une réflexion experte concernant les enjeux qui se posent aux *Official Statistics* – les statistiques officielles, ou publiques selon la terminologie communément retenue en français – en ce début de XXIe siècle. Il en propose une mise en perspective originale sous la forme d'un triptyque reliant présent, passé et futur, que portent les trois principales parties de l'opus. Le trajet suivi, se nourrissant de deux cents ans d'histoire mais tendu vers l'avenir, traduit une analyse volontiers prospective, engagée dans la promotion du rôle central des statistiques que l'auteur place au cœur du fonctionnement de la société.

Dans une courte partie introductive (I.), W. Radermacher pose les premiers jalons de définition des différentes facettes de son objet – les statistiques officielles – et fournit les clés de lecture de l'ouvrage : il y sera question d'institutions, de méthodes et de productions statistiques, facettes qui s'articulent les unes aux autres de façon spécifique au cours de l'histoire. S'appuyant notamment sur les travaux d'Alain Desrosières, W. Radermacher en délimite quatre périodes successives (*cf.* la figure 1.1, p. 4 et ci-contre) : de la phase 1.0 ayant vu la naissance des statistiques au début du XIXe siècle jusqu'à la phase 4.0 caractérisée par l'émergence des Big data, dans un parallèle suggestif avec les différentes ères techniques (de la première révolution industrielle à la dernière révolution numérique, qui est à l'œuvre depuis une dizaine d'années).

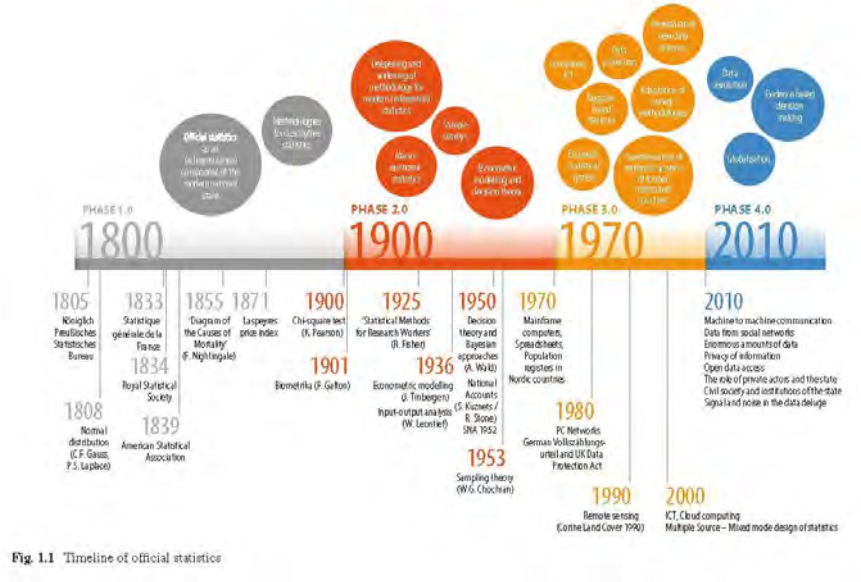


Fig. 1.1 Timeline of official statistics

La partie suivante (II.) revient en détail sur l'évolution des statistiques officielles des années 1970 aux années 2010, décennies qui correspondent à la phase 3.0 de la scansion historique proposée en introduction. Adoptant la terminologie des sciences de gestion, l'auteur décrit l'adaptation de leur « *business model* », qui suit désormais un processus industriel intégré, guidé par la recherche d'une qualité certifiée, allant de la conception des « produits statistiques » jusqu'à leur « consommation ». Ce n'est ainsi plus la myriade de productions artisanales fragmentées en autant d'approches thématiques, sectorielles ou nationales qui prévalaient auparavant. Selon W. Radermacher, il y a là un réel progrès de nature à renforcer la confiance des citoyens à l'égard des chiffres circulant dans le débat public. Coïncidant avec des avancées normatives – les principes fondamentaux des statistiques officielles, ainsi reconnues comme infrastructure publique d'information, ont été adoptés par l'ONU en 2014 (*cf.* l'encadré 2.3, p. 33-34) –, ce progrès s'appuie sur de nombreux facteurs : une main d'œuvre certes plus réduite mais nettement plus qualifiée ; une standardisation des méthodes de production, qui allient données d'enquête et registres administratifs ; une harmonisation des conventions de définition portée à l'échelle internationale ; une attention croissante à la confidentialité et à l'accessibilité des données ; et, de façon transversale, une meilleure capacité à s'adapter aux besoins de la société. Pour l'auteur, c'est grâce à ces évolutions que les statistiques officielles ont su répondre avec succès aux enjeux posés par les révolutions numériques successives et

des attentes politiques renouvelées, malgré des cadres budgétaires de plus en plus contraints et une distance croissante de la population vis-à-vis des dispositifs d'enquête.

Dans la troisième partie, intitulée *Science and society* (III.), le lecteur est embarqué dans une synthèse de réflexions épistémologiques concernant la place et le rôle des statistiques dans la société. D'utiles clarifications conceptuelles sont opérées, qui permettent à W. Radermacher de plaider pour une posture de réalisme critique s'écartant à la fois d'un réalisme naïf et d'un relativisme excessif. Cette posture est pour lui une condition nécessaire pour comprendre ce que sont véritablement les statistiques officielles : des représentations construites de la réalité, qui ne sont certes pas des vérités absolues mais n'en sont pas moins des informations de qualité auxquelles on peut (et doit) se fier. La notion de qualité est ici centrale, comme plus largement dans l'ouvrage. Et elle doit être entendue dans un sens large : elle ne renvoie pas seulement à des problèmes de méthodologie statistique, mais englobe les questions de gouvernance d'une part, de communication et d'appropriation par la société d'autre part. La réflexion proposée s'inscrit en cela dans le cadre conceptuel d'Alan Deming, spécialiste des organisations ayant inventé le *Total Quality Management*. Comme « objets frontière », fruits d'une co-construction de la science et de la société, les statistiques officielles sont également analysées sous un angle historique et sociologique. Ce sont des constructions politiques et des conventions sociales, et non de pures mesures de la réalité. Alors plus critique, l'auteur ne se limite pas à rendre compte de l'apport des statistiques. Il en souligne les excès dès lors qu'elles ne s'accompagnent pas d'une réflexivité suffisante et que le lien d'adhésion de la population se distend, parfois jusqu'à rompre, comme cela peut être le cas lorsque la société apparaît uniquement gouvernée par les nombres (p. 84 et suivantes). La partie s'achève par deux cas d'étude – les indicateurs et les statistiques du développement durable –, qui permettent d'illustrer les réflexions développées précédemment.

La dernière partie substantielle de l'ouvrage (IV.) ainsi que la très courte conclusion (V.) sont tournées vers le futur. W. Radermacher y souligne la nécessité d'un changement de paradigme des statistiques officielles en raison de l'ampleur des processus de globalisation et de numérisation en cours. Comme il l'indique, la numérisation est une réelle révolution : d'une part, avec la croissance exponentielle des puissances de calcul, de nouvelles données et potentialités d'analyse se font jour, qui remettent en question le monopole des statistiques officielles pour produire des informations quantifiées ; d'autre part, même si la demande de décisions fondées sur des diagnostics établis scientifiquement s'est imposée dans l'ère de quantification généralisée que nous connaissons, l'auteur invite à n'oublier ni les dangers liés au *quantitative turn*, où l'on accorde une attention bien trop grande à la quantité des données et bien trop faible à leur qualité, ni les risques qu'il entraîne par réaction, avec l'émergence de forces populistes promouvant des discours de post-vérité. La globalisation va quant à elle de pair avec le déclin des États nations, et donc des statistiques officielles qui leur restent fortement liées. Deux tendances opposées sont ainsi désormais à l'œuvre, qui posent toutes deux des difficultés spécifiques : vers le global, ce qui suppose que soient dépassés les blocages politiques pour élaborer des conventions d'équivalence internationalement partagées tout en veillant à ne pas fragiliser le lien de confiance existant avec les sociétés nationales ; et vers le local, ce qui implique de garder une indépendance suffisante vis-à-vis des pouvoirs publics, intérêts privés et des demandes de la population à l'échelle locale. Dans ce cadre, W. Radermacher espère que des solutions innovantes seront trouvées, qui associent nouvelles potentialités technologiques, transparence méthodologique, et garde-fous éthiques et juridiques. Il invite pour cela à lancer un débat sur le rôle des statistiques officielles dans la société, en plaçant au cœur des discussions les notions de co-production et de gouvernementalité afin que ce débat ne se limite pas à des considérations méthodologiques et intègre des réflexions issues des sciences sociales (*cf.* pages 126-127). Les thématiques de la communication et de la gouvernance devraient selon lui être centrales, puisqu'elles sont un moyen décisif de maintenir et même de renforcer le lien entre les statistiques officielles et les citoyens, qu'il s'agisse de professionnels amenés à jouer

un rôle nouveau dans leur production (comme les *data journalists* ou les *data scientists*) ou d'utilisateurs ordinaires de plus en plus en demande d'être associés. On ne peut que souscrire à un tel projet !

Avec ce livre, W. Radermacher propose un ensemble organisé de réflexions très bien informées sur l'évolution récente et à venir des statistiques officielles. Leur analyse formulée du point de vue des sciences économiques et de gestion – en termes de *business model*, produits, consommateurs ou encore d'avantages comparatifs – pourra surprendre un lecteur français habitué à penser ce domaine d'activité comme relevant *a priori* de l'État. Mais ce pas de côté n'est pas sans intérêt : il invite à réfléchir à la place et au rôle des statistiques officielles, comme lieu de rencontre de la science et de la société qui, s'il s'est constitué en premier sous l'égide de l'État, est amené à se réinventer face à la diffusion de modes de production et de régulation marchands dans le domaine de la connaissance comme du numérique. Au rang des critiques, on pourra regretter des développements parfois trop théoriques et abstraits (notamment sur la qualité et la gouvernance) et la place insuffisante accordée à des études de cas ancrées dans la réalité sociale. Ce choix éditorial peut paraitre en contradiction avec l'exigence pourtant bien formulée par l'auteur de voir les statistiques officielles accessibles à un public large. L'ouvrage n'est manifestement pas adressé à l'ensemble des « citoyens statisticiens » que W. Radermacher appelle de ses vœux. Cette critique n'ôte toutefois rien au grand intérêt que pourront y trouver les connaisseurs de la statistique publique et de son rôle dans la société, qu'ils soient eux-mêmes statisticiens (publics ou privés), responsables d'administration, chercheurs, journalistes ou même simples observateurs.

Le secret statistique

de Jean-Pierre LE GLÉAU (2019)

Gérard LANG[1]
Statisticien retraité, SFdS

Livre (199 pages)
Auteur : Jean-Pierre LE GLÉAU
Édition : EDP Sciences (Collection : Le monde des données) – 2019
ISBN : 978-2-7598-2332-1

1. Adresse mail ; Gérard Lang a fondé la division « Environnement juridique de la statistique » de l'INSEE en 1996 et a été secrétaire du comité du secret statistique de mai 1993 à septembre 2010.

L'éditeur EDP Sciences (« Édition Diffusion Presse Sciences »)[2] a publié en avril 2019 dans sa collection « Le monde des données » parrainée par la SFdS (« Société Française de Statistique ») un remarquable livre rédigé par Jean-Pierre Le Gléau[3], inspecteur général honoraire de l'INSEE, intitulé « Le secret statistique ».

Cet ouvrage de 199 pages comprend un sommaire (2 pages), une préface (11 pages), une introduction (5 pages), une première partie « Qu'est-ce que le secret statistique ? » (39 pages), une deuxième partie « Le secret statistique et la diffusion » (42 pages), une troisième partie « L'accès aux données confidentielles » (54 pages), une quatrième partie « Comment ça se passe ailleurs ? » (17 pages), une chronologie (3 pages) et une liste des principaux textes autour du secret statistique (13 pages), ainsi qu'un index (3 pages).

L'exposition est à la fois très claire, très précisément documentée et pratiquement exhaustive. C'est un véritable défi que d'y apporter quelques compléments d'information d'un certain intérêt, mais je ne me priverai pas d'essayer.

1. La préface

Le préfacier, Jean Gaeremynck, aujourd'hui président de la section des finances du Conseil d'État, a été président du Comité du secret statistique de 2009 à 2018 et cosignataire avec Maurice Méda en 1996 d'un rapport remis au Premier ministre proposant quelques assouplissements dans la loi de 1978 « Informatique et libertés ». C'est dire s'il connait aussi bien les aspects juridiques que pratiques du sujet, ce qui lui permet de signer un texte de très haute tenue qui l'amène notamment à s'interroger sur les questions de méthodologie liées à la participation de la statistique à l'évaluation des politiques publiques.

2. L'introduction

Dans son introduction, Jean-Pierre Le Gléau présente les enjeux qui sont liés à la protection des données individuelles dans une société moderne où la statistique publique se doit de fournir une information agrégée toujours plus pertinente et faisant l'objet d'une demande de plus en plus fine et détaillée. Il commente la position un peu avancée de la France résultant de l'adoption, notamment en réaction au projet de l'INSEE, nommé SAFARI (Système Automatisé pour les Fichiers Administratifs et le Répertoire d'Identification), d'informatisation du répertoire national d'identification des personnes physiques (RNIPP), de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés (qui n'a pas été sans créer des difficultés de compréhension mutuelle entre la Commission nationale de l'informatique et des libertés, CNIL, qu'elle a créée et l'INSEE à l'occasion des discussions sur les modalités de diffusion du recensement de 1982). La loi de 1978 a ensuite été modifiée par la loi n° 2018-493 du 20 juin 2018 modifiant la loi de 1978 pour tenir compte de l'entrée en vigueur au 25 mai 2018 du règlement européen (UE) 201/679 du 27 avril 2016 dit RGPD (Règlement Général sur la Protection des Données).

Jean-Pierre Le Gléau donne également quelques détails appétissants sur la nature particulière du secret en matière de statistique dans cette problématique générale de la protection des données individuelles et sur le rôle moteur joué par les règles applicables au secret statistique dans la mise en place de la confidentialité des données au sein de l'administration française.

2. Qui a été vendu le 28 juin 2019 à la société CSPM (« Chinese Science Publishing Media »).
3. Rappelons que Jean-Pierre Le Gléau a dirigé la rédaction du numéro spécial n° 63 de Statistique et Société, consacré à l'obligation de réponse.

3. La première partie : « Qu'est-ce que le secret statistique ? »

A – La partie « Qu'est-ce que le secret statistique ? » commence par exposer les fondamentaux de la loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques. Ce texte, qui constitue la charte fondamentale de la statistique publique française, est issu de la fusion voulue par le Conseil d'État de deux projets distincts relatifs, d'une part, aux statistiques d'entreprises et porté par le patronat et, d'autre part, aux statistiques sur les ménages et porté par l'Institut national de la statistique et des études économiques pour la métropole et la France d'outre-mer (INSEE) fondé en 1946 pour devenir la colonne vertébrale du système statistique public français.

Un encadré rappelle que le Conseil de la République (équivalent du Sénat sous la IVe République, mais ne s'exprimant qu'à titre consultatif) s'est opposé en 1951 très majoritairement et de manière virulente au principe même de la collecte d'informations relatives à la vie personnelle et familiale et, d'une manière générale aux faits et comportements d'ordre privé. Un orateur jugeait alors un tel projet « *absolument insoutenable, (et créant une) sorte d'inquisition nouvelle dans la vie personnelle, dans la vie privée, dans l'existence des personnes qui composent une famille et d'une façon générale dans le comportement intime de nos concitoyens* ». L'Assemblée nationale choisit de ne pas suivre l'avis du Conseil de la République et d'adopter pratiquement sans changement le texte présenté par le Gouvernement.

On comprend sans peine, dans ces conditions, la nécessité de l'équilibre qui apparait dans le titre même de la loi entre l'obligation, la coordination et le secret statistique. Si l'idée de la nécessité d'une coordination du système statistique public est née avec la création du Conseil supérieur de la statistique par un décret du 19 février 1885 publié au Journal officiel du 22 février 1885, où il est précédé d'un rapport d'une très grande hauteur de vue, celle de la nécessité d'une obligation de réponse (avec la sanction d'une amende en cas de résistance) n'est entrée dans le droit français qu'en 1938 par une série de trois décrets rédigés par Alfred Sauvy dans le cadre de la préparation d'une économie de guerre, et ne s'appliquait qu'aux statistiques d'entreprises (et une protection des données individuelles par le moyen du secret professionnel y figure).

La novation qu'instaure la loi de 1951 par la garantie donnée aux répondants que les informations individuelles qu'ils apportent à la statistique publique seront couvertes par un secret imperméable, beaucoup plus protecteur que la simple confidentialité traditionnelle attachée à toute information détenue par une administration, est donc un élément d'équilibre et de motivation parfaitement indispensable. Notamment, le secret statistique s'oppose frontalement au « droit général de communication » des administrations fiscales et douanières prévu par l'article L. 83 du livre des procédures fiscales s'appliquant à toute information légalement détenue par une administration ; et sa rupture est explicitement réprimée par le code pénal.

En ligne avec la réaction du Conseil de la République, la protection du secret statistique des données relatives aux ménages est quasi-absolue, alors que celle concernant les données relatives aux entreprises est exprimée de manière un peu plus relative.

L'unique dérogation générale au secret statistique découle de l'obligation faite aux fonctionnaires par l'article 40 du code de procédure pénale de dénoncer auprès du procureur de la République tout crime ou délit dont il acquiert connaissance dans l'exercice de ses fonctions.

La seule enquête statistique pratiquement susceptible de poser question dans ce cadre est l'enquête « cadre de vie et sécurité » (CVS), à laquelle est consacrée un encadré décrivant les

trésors de contorsion déployés par l'INSEE pour collecter les données de cette enquête sur micro-ordinateur en faisant tout pour éviter que l'agent enquêteur ne puisse en prendre lui-même connaissance, afin de ne pas avoir à dénoncer un éventuel crime ou délit qui apparaitrait dans la réponse au questionnaire.

B – Cette première partie se poursuit par une étude des modifications subies par la loi de 1951.

Ainsi, le secret statistique ne peut plus être présenté comme la contrepartie effective de l'obligation de réponse depuis qu'à partir de 2004 le programme statistique public annuel comprend à la fois des enquêtes obligatoires et d'autres qui ne le sont plus.

En outre, la durée de la protection des données par le secret statistique, conçue comme illimitée en 1951, a été ramenée à 100 ans pour ce qui concerne les données statistiques sur les personnes physiques par la loi n° 79-18 du 3 janvier 1979 sur les archives, puis à 30 ans pour ce qui concerne les données statistiques relatives aux entreprises par l'ordonnance n° 2004-280 du 25 mars 2004 relative aux simplifications en matière d'enquêtes statistiques.

Puis, la loi n° 2008-96 du15 juillet 2008 relative aux archives a encore réduit ces délais de protection à 75 ans pour les ménages et 25 ans pour les entreprises.

Une évolution très importante, résultant d'une demande de « régularisation de la situation » de la CNIL, a été apportée par la loi n° 86-1305 du 23 décembre 1986 portant modification de la loi du 7 juin 1951 qui autorise l'accès du système statistique public à toute donnée légalement détenue par l'administration (à l'exclusion de celles relatives à la santé et à la vie sexuelle des personnes physiques) et crée ainsi pratiquement un « droit de communication statistique ».

C – L'évolution la plus marquante concerne l'apparition progressive d'exceptions au secret statistique et la création au sein de la statistique publique d'une institution chargée d'organiser des dérogations de plus en plus importantes à ce secret, notamment en direction de la recherche.

L'ouverture sur cette question commence lorsque le patronat accepte la création dans le décret n° 84-628 du 17 juillet 1984 relatif au Conseil national de l'information statistique (CNIS) d'un comité du secret statistique concernant les entreprises, présidé par un conseiller d'État et auquel il participe, qui peut autoriser l'accès à des données d'entreprises couvertes par le secret statistique pour des finalités compatibles avec la lettre de la loi de 1951. Ainsi, le CNIS et le comité du secret statistique ont estimé en 1986 qu'il était possible de diffuser les effectifs globaux d'une entreprise et de chacun de ses établissements (une fois par an), ainsi que la catégorie d'importance de son chiffre d'affaires, la catégorie d'importance de son chiffre d'affaires réalisé à l'exportation et un indicateur de l'exercice (ou non) d'une activité de recherche en son sein.

Puis l'ordonnance n° 2004-280 du 25 mars 2004 a élargi l'ouverture initiale en élevant au niveau de la loi un nouveau « comité du secret statistique » dont les compétences sont élargies pour donner son avis sur les demandes de communication des données individuelles d'ordre économique et financier relatives aux personnes morales de droit public et de droit privé, et à l'activité professionnelle des entrepreneurs individuels et des personnes exerçant une profession libérale collectées en application de la loi de 1951.

La loi n° 2008-696 du 15 juillet 2008 a ensuite non seulement réduit les délais de protection du secret statistique, mais aussi ouvert la possibilité d'accès aux données individuelles relatives aux ménages en modifiant le régime de communication des archives publiques dans le sens d'une plus grande ouverture. À la suite, le décret n° 2009-328 du 20 mars 2009 relatif au

Conseil national de l'information statistique et au comité du secret statistique a pris acte de cet élargissement considérable des possibilités et réformé le comité du secret statistique en y créant deux sections, l'une compétente pour les renseignements individuels ayant trait à la vie personnelle et familiale et, d'une manière générale, aux faits et comportements d'ordre privé et l'autre compétente pour les renseignements individuels d'ordre économique ou financier.

4. La deuxième partie : « Le secret statistique et la diffusion »

A – Les questions relatives à la diffusion des données statistiques nécessitent de disposer d'une définition parfaitement limpide de la protection apportée par le secret statistique. Si la situation est assez simple concernant les données individuelles relatives aux entreprises (où, de plus, les identifiants SIRENE jouent un rôle très important), elle est compliquée dans le cas des données sur les personnes physiques par l'intervention des concepts introduits par les textes français et européens sur la protection des données individuelles. Un encadré « Risque ou aversion ? » analyse ce point, qui est ensuite illustré par un exemple.

B – Le secret statistique s'applique également dans le cas des données agrégées, notamment lorsque les agrégats concernés ne portent que sur un faible nombre d'individus. Les règles applicables à la diffusion de ces données, distinctes dans le cas des données sur les entreprises et des données sur les ménages, sont rappelées (avec un encadré sur les zones IRIS pour la diffusion des données du recensement général de la population de 1999).

C – Les règles relatives à la diffusion des données individuelles, également distinctes dans le cas des données sur les entreprises et des données sur les personnes physiques (selon que les fichiers soient considérés comme nominatifs ou non-nominatifs), sont également rappelées. Dans ce contexte, on trouve un développement très intéressant sur les notions de k-anonymat (voir les sources (32) et (41) de la documentation), de l-diversité et de t-proximité (voir la source (33) de la documentation) et les techniques utilisées à cet effet. La question des identifiants individuels SIRENE et surtout NIR (Numéro d'Inscription au Répertoire national d'identification des personnes physiques, RNIPP) est ensuite abordée, et mentionne l'assouplissement apporté par la loi n° 2016-1321 du 7 octobre 2016 pour une République numérique, s'agissant de l'utilisation du hachage et du recodage du NIR, d'une part, pour les traitements qui ont exclusivement des finalités de statistique publique et sont mis en œuvre par le système statistique public et, d'autre part, pour les traitements qui ont exclusivement des finalités de recherche scientifique ou historique.

D – La diffusion des données géographiques présente en France une difficulté particulière du fait du découpage administratif du territoire en près de 35 000 communes, dont la finesse n'a pas d'équivalent dans les autres pays européens. L'autre approche des données géographiques fines, basée sur la notion de carroyage, a également été expérimentée par l'INSEE et a donné lieu en 2013 à un regrettable incident, relaté dans un encadré.

E – Un élément supplémentaire important relatif à la diffusion des données statistiques a été introduit par la directive 2003/4/CE/du Parlement européen et du Conseil du 28 janvier 2003 concernant l'accès du public à l'information en matière d'environnement, transposée en droit français par la loi n° 2005-1319 du 26 octobre 2005 portant diverses dispositions d'adaptation au droit communautaire en matière d'environnement. La loi dit que l'autorité publique peut s'opposer au droit d'accès à l'information environnementale, s'agissant de données protégées par le secret statistique, après avoir mesuré les bienfaits comparés pouvant résulter, d'une part, d'un tel accès, et d'autre part, du respect de la confidentialité des données statistiques. Si la mise en œuvre d'un tel arbitrage semble devoir se poser rarement, je connais cependant un cas

de levée du secret statistique par le ministère de l'agriculture pour un motif d'urgence sanitaire et environnementale avérée, consistant à transférer la liste et la localisation des exploitations aviaires issues du recensement de l'agriculture aux services concernés dans le cadre de la lutte contre la grippe aviaire.

5. La troisième partie : « L'accès aux données confidentielles »

A – Initialement, l'accès aux données statistiques était réservé aux agents du service statistique public collecteur de ces informations. Il a ensuite rapidement été considéré comme légitime de pouvoir transférer de telles données à des agents d'un autre service statistique public. Les personnes concernées étaient tenues au secret professionnel et s'engageaient à ne rediffuser à quiconque aucune information couverte par le secret. Mais il est ensuite devenu clair que l'utilisation de ces données par des chercheurs, notamment dans le domaine économique et social, serait d'une grande utilité, à condition de mettre en place des modalités d'accès aux données concernées présentant suffisamment de garanties. Il est également clair que la question se pose différemment pour les données relatives aux entreprises et pour les données relatives aux personnes physiques.

B – L'auteur examine d'abord les conditions de l'accès aux données confidentielles relatives aux entreprises par l'entremise de l'une des quatre réunions annuelles du comité du secret statistique concernant les entreprises entre la création de celui-ci en 1984 et 2012. Les garanties apportées par les chercheurs concernant le respect de la confidentialité des données qui leur étaient transmises ne reposent à ce moment que sur la confiance mutuelle entre le chercheur et les membres du comité qui lui ont donné un avis favorable. En droit, cet avis favorable du comité ne lie pas les services enquêteurs concernés, qui pourraient refuser de donner suite à cet avis et de transmettre leurs données aux chercheurs habilités, mais cela ne s'est en fait jamais produit.

C – Vient ensuite l'examen des conditions d'accès aux données statistiques individuelles concernant les personnes physiques par l'entremise du comité du secret entre l'année 2009, où cet accès est devenu légalement possible, et 2012. Il n'était évidemment pas question pour la statistique publique d'adopter en la matière des procédures aussi libérales que celle alors en vigueur dans le cas des données d'entreprises. On a donc commencé par mettre en place une procédure de consultation sur place, dans les locaux du service statistique public détenant les données concernées, en faisant du chercheur qui avait convaincu le comité du secret de l'intérêt de sa recherche un agent provisoire de ce service statistique, « embauche à durée limitée » matérialisée par une convention entre le service statistique et l'organisme de recherche de rattachement du chercheur précisant notamment les conditions disciplinaires de l'exercice et l'obligation de respect du secret statistique.

Une autre modalité possible était le recours à un certain nombre de fichiers-détail de certaines enquêtes par sondage (dont l'enquête annuelle sur l'emploi) spécifiquement préparés par l'INSEE et accessibles par tout public sur Internet. Mais les exigences de la nécessaire anonymisation de ces fichiers dégradaient assez fortement l'information qui y restait disponible, et les rendaient le plus souvent « impropre à la consommation » par les chercheurs. L'Institut a donc décidé en 2006 de créer des « fichiers de production et de recherche » (FPR) spécialement adaptés aux besoins des chercheurs et apportant un plus grand niveau de détail que les fichiers précédents. Ces fichiers restaient anonymes en usage normal, mais une personne mal intentionnée et usant de données externes à ces fichiers aurait pu identifier un petit nombre des individus présents dans l'échantillon. On peut dire que ces fichiers n'étaient pas strictement anonymes, mais entraient dans le concept de fichiers « raisonnablement protégés » contre le risque d'identification des

personnes physiques y figurant, entrant dans le cadre de la législation européenne sur la protection des données individuelles, mais formellement « hors limite » dans la cadre d'un droit français encore plus strict. Et c'est pourquoi leur accès était très strictement contrôlé, les chercheurs habilités devant passer par l'intermédiaire de l'ADISP (Archives de Données Issues de la Statistique Publique), service inséré au sein du Centre Maurice-Halbwachs (CMH, ex-LASMAS : Laboratoire d'Analyse Secondaire et de Méthodes Appliquées à la Sociologie), unité mixte du Centre National de la Recherche Scientifique (CNRS), de l'École des Hautes Études en Sciences Sociales (EHESS) et de l'École Normale Supérieure (ENS). Depuis l'entrée en vigueur de la loi de 2016 sur la République numérique, l'habilitation d'accès à ces fichiers bénéficie d'une procédure simplifiée devant le comité du secret statistique.

D – Depuis 2012, la mise en place d'un « Centre d'Accès Sécurisé aux Données » (CASD) au sein du Groupe des écoles nationales de statistique et d'économie (GENES, qui était initialement l'une des huit directions faisant partie des services centraux de l'INSEE, avant de devenir à compter du 1er janvier 2011 un établissement public à caractère scientifique, culturel et professionnel placé sous la tutelle technique de l'INSEE) a profondément modifié les conditions d'accès des chercheurs aux données statistiques individuelles, s'agissant des données d'entreprises comme des données relatives aux ménages, et apporté des garanties de protection du secret statistique satisfaisant les critères les plus exigeants en la matière. Un encadré détaille précisément les techniques mises en œuvre au sein du CASD et les garanties de sécurité résultant de l'utilisation de la « SD-Box » dans ce cadre.

Un arrêté du 20 décembre 2018 a ensuite transformé le CASD en un Groupement d'Intérêt Public (GIP) associant cinq partenaires (L'État, représenté par le ministre chargé de l'économie, lui-même représenté par le directeur général de l'INSEE ; le GENES ; le CNRS ; l'École polytechnique ; HEC Paris, Hautes Études Commerciales).

E – Après une longue résistance des services fiscaux (illustrée par un encadré intitulé « Une histoire de virgule »), les données fiscales (protégées par le secret fiscal défini par l'article L. 103 du code des procédures fiscales) sont devenues accessibles aux chercheurs par l'entremise du comité du secret statistique et par le moyen du CASD, aux termes de l'article 104 de la loi n° 2013-660 du 22 juillet 2013 relative à l'enseignement supérieur et à la recherche.

F – Les données individuelles relatives à la santé des personnes physiques sont parmi les plus sensibles et leur protection fait donc l'objet d'un renforcement particulier dans la loi « Informatique et libertés ». Aussi l'accès à ces données était avant 2016 d'une grande complexité. L'article 193 de la loi n° 2016-41 du 26 janvier 2016 de modernisation de notre système de santé (qui fait l'objet d'un encadré spécifique), tel que (heureusement !) rectifié par l'article 37 de la loi sur la République numérique, puis encore modifié par le nouveau chapitre IX (« Traitements de données à caractère personnel dans le domaine de la santé ») résultant de l'article 16 de la loi n° 2018-493 du 20 juin 2018 relative à la protection des données personnelles, de la loi de 1978 ensuite réécrite par l'ordonnance n° 2018-1125 du 12 décembre 2018, puis par l'article 41 de la loi n° 2019-774 du 19 juillet 2019 relative à l'organisation et à la transformation de notre système de santé, a revu en profondeur le dispositif d'accès à ces données. Le nouveau dispositif comprend le comité de protection des personnes (CPP) pour les demandes d'autorisation relatives aux recherches portant sur la personne humaine et le comité d'expertise pour les recherches, les études et les évaluations dans le domaine de la santé (CEREES). On doit cependant déplorer que le ministère de la santé ait choisi de pérenniser un dispositif spécifique complètement séparé en substituant à l'Institut national des données de santé, (NDS) une « Plateforme des données de santé » (Health data hub, HDH, pour faire moderne), constitué sous la forme d'un Groupement d'intérêt public rassemblant huit ministères et une cinquantaine des personnes morales de droit public et de droit privé (dont la convention de constitution a été approuvée par un arrêté du 29 novembre 2019, et dont le

CASD ne fait pas partie) pour donner accès à des données, ce qui ne permet pas de les croiser avec les données individuelles accessibles par le canal du CASD

G – Certaines données individuelles relatives aux entreprises détenues par la Banque de France (BDF) ont un statut particulier, dans la mesure où la BDF fait partie du Système européen des banques centrales (SEBC). Ainsi, l'accès aux données confidentielles transmises à la BDF du fait de son appartenance au SEBC est régi par le règlement (CE) n° 2533/98 du Parlement européen et du Conseil du 23 novembre 1998 concernant la collecte d'informations statistiques par la Banque centrale européenne (BCE). L'habilitation d'accès à ces données est délivrée par le « Comité d'examen des demandes d'accès aux données de la Banque de France » et les données sont uniquement accessibles dans une « Open Data Room » située dans les locaux de la BDF. De ce fait, il est également impossible de croiser les données de la Banque de France avec d'autres données accessibles au CASD, ce qui peut être dommageable pour certaines recherches.

H – En outre, l'article L. 311-8 du code des relations entre le public et l'administration (CRPA), créé par l'article 36 de la loi n° 2016-1321 du 7 octobre 2016 pour une République numérique (et ultérieurement marginalement modifié par l'article 4 de la loi n° 2018-670 du 30 juillet 2018 relative à la protection du secret des affaires), qui fait l'objet d'un encadré intitulé « un article à la rédaction et à l'utilité contestables », élargit encore le champ de compétence du comité du secret statistique. Cet article prévoit que lorsqu'une demande de dérogation aux délais de protection du droit du code du patrimoine vise à effectuer des traitements à des fins de recherche ou d'étude présentant un caractère d'intérêt public à partir d'une base de données, l'administration qui détient la base de données concernée ou l'administration des archives peut demander l'avis du comité du secret statistique.

6. La quatrième partie : « Comment ça se passe ailleurs ? »

A – Cette partie donne des informations décrivant la situation du secret statistique aux Nations Unies, dans l'Union européenne et dans les pays suivants : le Canada, les Pays-Bas, le Royaume-Uni, la Russie et la Tunisie (qui présentent des cas assez différents).

B – J'ajoute ici une note sur un autre cas intéressant, la Suisse, pays voisin du notre mais qui n'est pas partie au traité sur l'Espace économique européen (dit « EEE »). De ce fait, la Suisse ne participe pas au « système statistique européen étendu aux pays membres de l'association européenne de libre-échange (AELE : Norvège, Islande, Liechtenstein) » et pratique un secret statistique fort rigide, qui s'oppose strictement à tout transfert de données individuelles à l'extérieur du pays. Ainsi, il n'est pas possible d'échanger des données relatives aux travailleurs français frontaliers avec la Suisse.

Paradoxalement, la Suisse est le seul pays pour lequel j'ai connaissance d'une violation grave et avérée du secret statistique, à la suite du transfert volontaire d'un fichier important de données statistiques individuelles à l'extérieur des bureaux de l'Office fédéral suisse de statistiques de Neuchâtel. Le responsable de cet acte a été poursuivi et extradé d'Espagne, où il avait pensé pouvoir se réfugier. Je précise également qu'en ma qualité de secrétaire du comité du secret statistique, j'ai eu l'occasion d'exercer des représailles contre la dureté du secret statistique suisse en m'opposant, au nom de la réciprocité, au transfert de données statistiques individuelles françaises à une chercheuse de nationalité française exerçant habituellement ses talents à l'Université d'Aix-Marseille (où elle les aurait obtenues sans problème), mais qui avait le malheur de les demander dans le cadre d'une recherche s'effectuant à Genève !

7. La chronologie

La chronologie retrace les principales étapes marquant le développement de la problématique du secret statistique depuis la création de l'Institut national de la statistique et des études économiques pour la métropole et la France d'outre-mer (INSEE) par les articles 32 et 33 de la loi de finances du 27 avril 1946 jusqu'à un arrêté du 20 décembre 2018 portant approbation de la convention constitutive du groupement d'intérêt public « Centre d'accès sécurisé aux données », qui acte la transformation du CASD, initialement créé au sein du Groupe des écoles nationales d'économie et statistique (GENES) en un GIP rassemblant cinq partenaires.

8. La liste des principaux textes autour du secret statistique

La liste des principaux textes autour du secret statistique dont les (62) références sont rassemblées dans la partie 6 du livre est particulièrement fouillée. Tout au plus peut-on éventuellement regretter l'absence de référence à deux textes supplémentaires d'un certain intérêt :

- La déclaration sur l'éthique professionnelle adoptée le 23 août 1965 par résolution de l'assemblée générale de l'Institut international de statistique (IIS/ISI).
- La recommandation n° (97) 18 adoptée le 30 septembre 1997 par le Comité des ministres du Conseil de l'Europe concernant la protection des données à caractère personnel collectées et traitées à des fins statistiques.

9. L'index

L'index est très utile. J'aurais simplement aimé y trouver l'entrée « RNIPP : Répertoire national d'identification des personnes physiques ».

Je signale également qu'aucune mention n'apparait dans le texte (et donc dans l'index) concernant le répertoire électoral unique (REU, créé par la loi n° 2016-1048 du 1er août 2016), qui est certainement l'un des fichiers les plus importants gérés par l'INSEE.

Pour les puristes, remarquons que le site francophone de l'ONU écrit « Organisation des Nations Unies », ce qui n'est ni la forme figurant dans le texte de la quatrième partie, ni celle reprise dans la référence (43) de la partie 6, ni celle figurant dans l'index !

Enfin, confions notre soulagement de constater l'absence de toute mention de « big data », « data management » ou « data scientist » dans le texte, qui n'offre qu'une seule occurrence pour « clef USB », pour « CD-ROM » et pour l'« Open Data Room » de la BDF et, surtout, mentionne, dans l'encadré très technique de la troisième partie sur le « CASD du GENES » (et dans l'index) la précieuse « SD-Box » (But what does it mean ?).

10. Pour conclure

Au total, et même si la matière juridique qu'il aborde peut parfois être considérée comme localement ardue, voilà un ouvrage splendide, voire incontournable, à mettre dans les mains de tout statisticien (au sens large) soucieux de comprendre les relations entre son activité professionnelle (même privée) et les exigences et perspectives de la société de l'information dans laquelle nous baignons aujourd'hui.