Préface à la deuxième édition

Statistique en action : ainsi s'appelait la première édition de cet ouvrage. Cette deuxième édition en constitue une version légèrement revue et surtout augmentée [1] dans son contenu imprimé. En effet, l'essentiel des plus de deux cents pages d'annexes disponibles en ligne [2] accompagnant la première édition (soit, trois corrigés et deux chapitres de compléments de cours) a pu être intégré dans cette édition. Ainsi, tous les problèmes, sauf le dernier, trouvent dorénavant leur corrigé dans la version imprimée du manuel.

Les retours que nous avons eus, essentiellement de la part de collègues enseignant la statistique mathématique mais aussi de quelques agrégatifs, nous ont confortés dans les partis pris et les choix qui nous avaient guidés lors de la rédaction de ce livre. Nous avions notamment décidé de présenter un cours complet mais condensé, allant droit à l'essentiel. Ce dernier a ainsi pu être utilisé comme trame pour élaborer un cours plus personnel, adapté aux goûts de chacun. Des lecteurs scrupuleux, Gérard Biau, Lucien Birgé et Jean-Michel Marin pour ne pas les citer, nous ont apporté des critiques constructives et leur avis a beaucoup compté dans les légères modifications apportées dans la première partie de cet ouvrage. Nous souhaitons les remercier chaleureusement.

La motivation initiale pour rédiger les textes illustratifs et leurs corrigés était de préparer les étudiants à l'oral de modélisation de l'agrégation. Nous avons constaté avec plaisir que nombre de ces problèmes étaient effectivement donnés comme planches de préparation à l'oral de l'agrégation. Il n'en demeure pas moins, comme il est expliqué dans l'avant-propos, qu'ils peuvent être employés à de nombreuses autres fins : pour illustrer et mettre en action le cours condensé (qui est aride à dessein), pour enrichir ses horizons statistiques et lire des introductions à divers domaines de recherche... ou, de manière plus utilitaire, en tant que sujets de mémoire de licence 3 ou de master 1. Les agrégatifs qui découvrent le présent ouvrage lors de leur année de préparation n'ont généralement pas le temps de lire en profondeur et de méditer tous les corrigés : nous ne pouvons donc que conseiller aux étudiants de master 1 d'anticiper leur préparation éventuelle de l'agrégation (et/ou leur poursuite en master 2 recherche) et d'utiliser ce manuel le plus tôt possible, dès leur premier cours de statistique mathématique.

[1] Et même dans son titre, qui est désormais *Statistique mathématique en action*

[2] Il y reste encore quelques compléments techniques, essentiellement de niveau master 2, à l'adresse `http://www.math.ens.fr/statenaction`

Nous voudrions conclure cette préface par des remerciements à l'attention de notre éditeur : nous mesurons la confiance qu'il nous accorde en nous proposant de publier cette version imprimée largement augmentée.

Paris, novembre 2011

Avant-propos : Statistique en action

La statistique semble souvent, même aux yeux des probabilistes, une matière mathématique mystérieuse. Les tenants et aboutissants de la démarche statistique, son vocabulaire, son appel à des choix nécessaires et parfois subjectifs dans le traitement des données déroutent. Le monde serait plus simple s'il se dégageait une vision claire de tout phénomène étudié, qui permettrait de le décrire parfaitement par un modèle mathématique. Malheureusement, c'est rarement le cas : l'adéquation entre le modèle proposé et les données recueillies lors de l'observation du phénomène est très souvent sujette à caution. Lorsqu'est proposé un modèle mathématique où le hasard est prépondérant, un des rôles éminents de la statistique mathématique est de quantifier l'accord entre les données disponibles et ce modèle. À cet égard, la statistique constitue en quelque sorte la physique des nombres. Elle a également pour objet de construire des procédures d'estimation ainsi que des méthodes d'encadrement pour mesurer leur précision. C'est, par exemple, le principe des fourchettes associées aux fréquences indiquées par les instituts de sondage. Pour paraphraser la définition « Majorer – Minorer – Approcher » que donne Dieudonné au début de son traité d'analyse, on verra dans le cours et les textes qui constituent ce manuel que la statistique reflète bien l'adage : « Modéliser – Estimer – Tester ».

L'idée de ce livre a germé au contact de nos étudiants mathématiciens et de nos collègues chercheurs. Plus ou moins familiers du calcul des probabilités, ces derniers sont pourtant souvent troublés par la statistique et ont même du mal à en saisir l'objet. Pour répondre à leurs attentes et interrogations, nous avons coutume d'expliquer que la statistique, comme les probabilités, joue avec le hasard, mais que l'objectif du statisticien est symétrique de celui du probabiliste. Ce dernier cherche à exhiber les propriétés d'une variable aléatoire dont il connaît (à des degrés divers) la loi, alors que le statisticien part des observations d'une variable aléatoire pour en déduire les caractéristiques de la loi dont elle est issue. Cette mise en perspective de la statistique, qui illustre ses liens profonds avec les probabilités par un nécessaire jeu de va-et-vient, montre qu'elle ne peut toutefois être vue comme une branche de ces dernières. C'est pourquoi un autre point de vue et des ouvrages spécifiques sont nécessaires. Il existe déjà d'excellents manuels, très divers, qui témoignent de la grande variété des approches de la statistique que l'on peut adopter. Beaucoup respectent la forme d'un manuel mathématique classique : énoncés de concepts et de résultats avec leurs démonstrations,

suivis de courts exemples. D'autres sont davantage centrés sur l'application concrète des formules ou des techniques statistiques que sur les éléments de preuve mathématique.

L'originalité de cet ouvrage se situe dans la *mise en action* de la démarche et des notions fondamentales présentées dans une première partie qui constitue un cours condensé relativement classique de statistique mathématique de niveau master. La seconde partie de cet ouvrage propose alors, sous forme de textes (et de leurs corrigés), une introduction succincte à un grand nombre de domaines de la statistique. Dans ces textes, nous ne nous contentons pas d'illustrer les notions du cours par des exemples. Ces dernières y sont rendues vivantes et mises en perspective. Ainsi, le premier pas de l'approche statistique est de construire un modèle : en se confrontant à ces textes, le lecteur pourra appréhender différents modèles statistiques (paramétriques ou non-paramétriques) et les appliquer à la détection du hasard dans les jeux de grattage, à la classification automatique ou à la compression de données. Le second pas de la démarche est souvent d'estimer les paramètres d'intérêt : en considérant de nombreux exemples d'estimateurs, et pas seulement une définition générale, le lecteur se familiarisera avec l'intuition et la créativité situées au fondement de cette étape cruciale de la résolution d'un problème de statistique. Le troisième et dernier pas, qui intéresse certainement les mathématiciens, est de quantifier la qualité de ces estimateurs : à cet effet, sont par exemple discutées dans les textes les notions de consistance, de normalité asymptotique ou de risque intégré. Dresser la liste des qualités que peut avoir un estimateur aboutirait à un inventaire à la Prévert : dans chaque texte défauts et qualités des estimateurs sont soulignés en fonction du contexte.

À qui nous adressons-nous ?

Aux agrégatifs, certainement (tant qu'il y en aura). Une des lames de fond ayant mené à ce livre est la réforme du concours de l'agrégation de mathématiques, qui a introduit en 1999 une épreuve de modélisation à l'oral. Cette dernière repose sur l'étude d'un texte, à partir duquel il est demandé à l'étudiant de mettre en application ses connaissances et son savoir-faire afin de discuter la modélisation d'un problème concret, résoudre les questions mathématiques inhérentes à cette modélisation, et, enfin, en illustrer informatiquement les résultats. C'est bien ce même principe que nous avons suivi en rédigeant la seconde partie de cet ouvrage. Les agrégatifs pourront utiliser nos textes de mise en action statistique comme autant d'occasions de s'entraîner sur des planches d'oral. À cet égard, nous avons essayé de fournir quelques textes sur des thèmes classiques (estimation non-paramétrique d'une fonction de régression ou d'une densité de probabilité, classification, estimation dans un modèle avec censure). Nous proposons également d'autres textes sur des sujets peut-être plus originaux (compression de données ou bandits à deux bras) mais pour qui les outils mis en œuvre entrent pleinement dans le programme exigible de l'agrégation. Bien entendu, tout le champ de la statistique n'est pas représenté et on pourra toujours déplorer telle ou telle absence. Cela reflète un certain parti pris mais également notre propre culture du domaine.

Les textes sont de longueurs et de difficultés variables, et certains ne pourraient pas vraiment être traités en temps limité pour un oral blanc. Nous avons pu le constater dans les lieux (École normale supérieure, Université Paris-Sud) où les étudiants s'y sont confrontés. Nous remercions au passage ces étudiants (des promotions 2003–04 à 2008–09) pour leurs critiques et encouragements ! La rédaction des textes varie : de la modélisation introduite de manière parfois succincte (auquel cas le corrigé, lui, détaille ce que le texte a omis) aux introductions plus bavardes (que le corrigé, là aussi, veille à améliorer en les formalisant). Si ces textes sont parfois imparfaits à dessein, nous avons en revanche porté le plus grand soin aux corrigés, qui comportent en particulier des illustrations informatiques et sont suivis ou précédés d'une liste de questions qu'un jury d'agrégation aurait pu poser. Ils font également mention d'une bibliographie. Par ailleurs, certains d'entre eux proposent également l'utilisation de jeux de données réelles (tombés dans le domaine public) afin d'aller au bout de la démarche qui consiste à confronter les résultats théoriques de la statistique mathématique à la réalité.

L'étude d'un document ne peut se faire sans connaître les résultats essentiels du domaine. Les agrégatifs qui cherchent un cours complet et condensé de statistique le trouveront dans la première partie de cet ouvrage. Celle-ci consiste en un énoncé de tous les résultats fondamentaux, ainsi qu'en une introduction de techniques plus avancées (par exemple la méthode delta) et d'objets (la p–valeur) auxquels les candidats n'auront pas nécessairement été confrontés au cours de leur préparation académique. Quelques preuves sont détaillées – en particulier lorsqu'elles sont originales –, la plupart sont simplement esquissées, quelques-unes sont omises, car trop longues ou fastidieuses, mais dans tous les cas, des références précises aux ouvrages que nous apprécions sont fournies. Comme toujours, il est essentiel que les agrégatifs ouvrent de nombreux livres pour mieux s'imprégner de la discipline.

Le cours condensé et les textes de mise en action s'adressent également à d'autres publics. Les étudiants de master 2 de probabilités et statistique pourront apprécier l'exposé synthétique des résultats qu'ils doivent connaître avant de commencer leur année, et trouveront dans les textes autant d'introductions à des cours qu'ils pourront suivre durant leur formation. C'est à eux que s'adresse plus particulièrement la bibliographie mentionnée à la suite de chaque corrigé. Les thèmes proposés fonctionnent comme autant d'enrichissements culturels nécessaires, car la tendance actuelle du système de recherche pousse à une grande spécialisation qui, parfois, va à l'encontre de l'acquisition d'une vue d'ensemble sur sa discipline. Bien mieux, les étudiants de master 2 qui se destinent à une carrière de chercheur trouveront dans les textes proposés certains outils fondamentaux d'axes de recherche modernes qui s'inscrivent parmi les plus féconds de la statistique d'aujourd'hui.

Les étudiants de master 1 pourront suivre notre cours synthétique au fur et à mesure de leur semestre de statistique. Ils pourront lire les textes pour s'imprégner des notions et de la démarche. Bien sûr, certains textes leur sembleront difficiles. Il n'en demeure pas moins que nous espérons très sincèrement que cet ouvrage leur procurera le « déclic statistique ». Déclic statistique ? Nous avons remarqué, à la

fois sur nous-mêmes et sur nos étudiants, et ce quel que soit l'enseignant, que les démarrages en statistique sont difficiles pour les étudiants qui ont uniquement suivi un cours de probabilités. Certaines assertions, comme

L'estimateur $\hat{\theta}_n$ converge alors vers le vrai paramètre θ.

sont parfois mystérieuses et peuvent ne trouver un sens que plus tard, généralement lors d'un deuxième cours de statistique, lors de la préparation à l'agrégation ou en master 2. Dans nos corrigés de textes, de telles assertions sont présentes à foison. Nous espérons que les voir dans un contexte réel aidera à ce déclic. Enfin, plus prosaïquement, de nombreux textes peuvent constituer des sujets de mémoire. Ce qui nous entraîne vers la dernière catégorie de mathématiciens visés...

... Nos collègues : ils ont parfois le sentiment d'avoir abordé de nombreux domaines des mathématiques (de la théorie des nombres à la géométrie, en passant par l'analyse harmonique, les EDPs et les probabilités), à l'exception de la statistique. Dans le système universitaire français, cette dernière est rattachée aux mathématiques. Dans la plupart des pays occidentaux, et notamment aux États-Unis, les départements de mathématiques et ceux de statistique coexistent. Le clivage est alors encore plus profond. En France, la statistique a la chance d'être reconnue par les mathématiciens ; reconnue mais pas toujours connue. Nous dédions donc ce livre à tous nos collègues qui nous ont un jour demandé une référence pour débuter en statistique ou qui ont suivi un de nos exposés de recherche sans même en comprendre la démarche. À tous ceux-là, le cours que nous proposons formera une introduction aux grands résultats culturels de la statistique.

Remerciements

Cet avant-propos ne saurait être conclu sans un hommage plus personnel à tous nos collègues préparant ou ayant préparé à l'agrégation : nous avons beaucoup communiqué entre nous, parfois par pages web personnelles interposées. Merci à Fabienne Castell, Djalil Chafaï, Cécile Durot, Aurélien Garivier, Sophie Lemaire, Christophe Leuridan, Florent Malrieu, Alain Trouvé et Mathilde Weill. Grâce à vous, nous avons pu lire de nombreux textes intéressants sur des domaines variés. Nous n'avons pas manqué de consulter les spécialistes de certains domaines traités dans ce livre : Bernard Bercu, Agathe Guilloux, Gábor Lugosi et Patricia Reynaud-Bouret, qui nous ont guidés ou confortés dans nos choix. Enfin, nous souhaitons remercier tout particulièrement nos relecteurs statisticiens ou probabilistes : Sylvain Arlot, Florent Autin, Jérémie Bettinelli, Gérard Biau, Dominique Bontemps, Sébastien Bubeck, Aurélien Garivier, Sébastien Gerchinovitz, Erwan Le Pennec, Thanh Mai Pham Ngoc, Marie Théret et Amandine Véber.

Paris – Orsay, septembre 2009