

Le titre de cette rubrique reprend volontairement bien sûr le titre de l'ouvrage mais également l'adresse web de son complément en ligne. En effet, les deux forment un document représentant plus de 500 pages. Le livre s'ouvre sur un avant-propos de quelques pages destiné essentiellement à présenter les motivations, les objectifs et les choix adoptés pour son contenu. En dehors d'être très convainquant, il est surtout conforme au contenu de l'ouvrage. Tout d'abord, le public premier du livre est celui des agrégatifs dans le but de préparer l'épreuve orale de modélisation. Pour cela, les auteurs proposent un cours condensé de statistique mathématique de niveau master 1/2 suivi d'une mise en situation sur huit textes d'application. Chaque sujet est abordé à travers une étape de modélisation statistique, puis un traitement critique du problème de l'estimation des « paramètres » du modèle. Une liste de suggestions de développements est également incluse. Chaque texte de mise en situation est accompagné d'une solution détaillée, de questions potentielles posées par un jury, de conseils pour le candidat/étudiant et d'illustrations numériques. Le code Matlab complet est fourni sur la page web des auteurs. Le matériel fourni s'appuie sur une forte expérience des auteurs dans la préparation à l'agrégation.

Cet ouvrage sera une référence indispensable à toute personne préparant l'agrégation mais également une source (d'inspiration) de projets pour des étudiants de master ou d'écoles d'ingénieurs. Pour la communauté mathématique, sa lecture est un très bon moyen de découvrir les statistiques et de voir comment différents outils standard de mathématiques appliquées peuvent intervenir pour résoudre un problème de statistique. En particulier, cela montre qu'une formation solide en statistique ne peut pas être déconnectée d'une solide culture mathématique. S'il restait encore des indécis, on peut mettre en avant que l'investissement réalisé pour l'achat de cet ouvrage est « minime » au vu du matériel fourni par les deux

auteurs !

Pour terminer, nous reprenons la table des matières de l'ouvrage et de son complément en ligne avec quelques commentaires.

## CONTENU DU LIVRE (287 PAGES HORS BIBLIOGRAPHIE ET INDEX)

### **Cours de statistique : synthèse et mise en perspective (105p)**

**Chapitre 1-9 Concepts fondamentaux de la statistique.** Estimation. Intervalles et régions de confiance. Tests d'hypothèses. Vecteurs gaussiens. Tests du  $\chi^2$ . Modèle linéaire gaussien. Fonctions de répartition. Simulation d'échantillons de loi donnée et applications

Des éléments classiques de statistique mais présentés avec un effort certain de synthèse. Quelques pistes de prolongements sont également mentionnées. On peut tout de même regretter l'absence d'un chapitre sur la statistique bayésienne. Notons une introduction de la notion de région de confiance avant celle de test, fondée d'après les auteurs, et à raison d'après ma propre expérience, sur une plus grande facilité des étudiants à appréhender le premier concept que le second. Une seconde « originalité » par rapport à la littérature statistique française est l'introduction de la notion de  $p$ -valeur. Il s'agit là encore, à mon avis, d'une excellente initiative.

### **Mise en action : 8 thèmes de statistique**

#### **Chapitre 10 : Machines à sous**

Problème de définition d'une stratégie optimale d'un joueur face à un « bandit » à deux bras, chaque bras admettant une loi de gain de type Bernoulli. Utilisation de techniques martingales. Ce texte peut être vu comme une incursion dans le domaine de l'apprentissage par renforcement pour aller, par exemple, vers des modèles avec des bandits multi-bras ou des modèles de type processus markoviens de décision (MDP).

#### **Chapitre 11 : Estimation non paramétrique pour le modèle de régression**

Un thème classique avec l'estimation d'une fonction de régression, ici, une fonction de  $L^2([0, 1])$  via une représentation de Fourier. Étude du risque quadratique de l'estimateur à l'aide de propriétés de régularité de la fonction de régression introduite via son appartenance à des classes d'espace de Sobolev. Le contrôle de risque est proposé en des termes adaptés à des connaissances de niveau master. Une discussion du choix du paramètre de seuillage dans la représentation de Fourier est également incluse.

#### **Chapitre 12 : Inférence statistique pour des modèles censurés**

Estimation non-paramétrique de la fonction de survie à partir de données indépendantes de durées de vie censurées (ou non). L'étude se concentre sur l'estimateur de Kaplan-Meier (l'estimation de la fonction taux de hasard n'est pas discutée). Ce thème peut être vu comme une introduction à l'analyse statistique de données sur un événement non-récurrent issu de nombreux domaines d'application : études biomédicales, fiabilité des systèmes, sciences actuarielles...

. De plus, il s'agit d'une brique de base dans l'analyse de phénomènes récurrents avec une modélisation paramétrique, non-paramétrique ou semi-paramétrique par processus ponctuels (modèle à risque proportionnels, modèle de Cox),

### **Chapitre 13 : Étude du nombre de renouvellements**

Cette partie peut être vue comme un complément au thème précédent, dans le sens où il s'agit d'analyser des données indépendantes (et de même loi) sur un événement récurrent. Le texte fournit des éléments sur le contrôle du nombre de renouvellements observés sur un intervalle de temps (en particulier, dans le cas d'un modèle de durée de vie de type New Better than Used). Ce thème est plus important qu'il n'y paraît, car l'étude du modèle de renouvellement est un premier pas vers l'analyse, en particulier asymptotique, de modèles tels que les processus de Markov, semi-Markov, régénératifs, au coeur de nombreux domaines d'applications : évaluation des performances de systèmes, bio-statistique,

### **Chapitre 14 : Estimation de densité de probabilité**

Thème classique de l'estimation non-paramétrique de la densité d'une loi d'un échantillon. Dans un premier temps, la consistance et la normalité asymptotique des estimateurs par histogramme et par fenêtres glissantes sont étudiées. Dans un second temps, la consistance d'estimateurs à noyaux est traitée. Notons que les auteurs proposent l'utilisation de la distance en variation totale comme critère d'erreur. Ainsi, la consistance est ici relative à la convergence en probabilité d'un critère global de type  $L^1$  entre l'estimateur et la densité cible. Dans le cas des estimateurs à noyaux, cela requiert de faire largement appel à la notion de produit de convolution. Cette partie se conclut par une application à la reconnaissance des formes.

### **Chapitre 15 : Classification de données**

Un thème central de statistique en vue des applications : classer des objets en faisant le moins d'erreur possible. Ici, les auteurs se placent dans un cadre de classification binaire. La fonction de régression issue du classifieur de Bayes est analysée par des techniques d'histogrammes. Une estimation par noyaux dans l'esprit du précédent thème est également abordée.

### **Chapitre 16 : Compression de données**

Un thème à l'interface entre probabilité, statistique et théorie de l'information. Problème de codage de mots aléatoires avec un loi d'émission inconnue, un des objectifs étant qu'il n'y ait pas de perte d'information. Introduction de la notion d'entropie d'une loi de probabilité et de ses relations avec la divergence de Kullback-Leibler. Les codages de Shannon, arithmétique et de Huffman sont abordés. Ce thème peut apparaître comme à la frontière de la « statistique traditionnelle », mais en termes de potentiel d'application, les enjeux sont considérables.

### **Chapitre 17 : Jeux de grattage**

Construire un test statistique de la répartition totalement au hasard des tickets donnant un gain significatif dans la confection d'un lot de tickets. Diverses variations sont proposées en fonction du niveau d'approximation de certaines lois de probabilité mise en jeu (en particulier des lois associées à des modèles de comptage). À partir du rejet d'une telle répartition totalement au hasard, il est proposé une discussion sur l'existence de stratégies de gain pour un joueur « informé ».

**Commentaires sur cette partie.** Un espace limité commande de faire des choix sur les thèmes abordés. On retrouve ici des sujets incontournables, par exemple 11,12,14,15. On pourra toujours regretter l'absence de texte centré, par exemple, sur la modélisation statistique à variables latentes classique (comme les modèles markoviens cachés) ou sur la statistique bayésienne, qui connaissent de nombreuses applications comme dans le domaine du traitement du signal, de la bio-statistique... On aurait pu imaginer également une introduction aux procédures de tests de comparaison multiple

et le contrôle de mesures de risque adaptées aux données de grande dimension.  
Au vu du travail réalisé, au soin apportée aux solutions commentées des 8 thèmes proposés ici, on ne peut qu'encourager les auteurs à préparer un autre volume de mise en action de la statistique !

## **Corrigés des thèmes de statistique en action**

**Chapitre 18-21** Corrigés des thèmes 11, 12, 15, 16 (125 pages)

### COMPLÉMENTS EN LIGNE DU LIVRE (224 PAGES)

**Chapitre 22-25** Corrigés des thèmes 10, 13, 14, 17 (126 pages)

## **Compléments techniques**

**Chapitre 26-27** Des rappels sur la théorie de l'intégration et des probabilités ; puis des compléments sur l'estimation (13 pages).

**Chapitre 28-35 : Des compléments sur les 8 thèmes de statistique en action** (80 pages) Ces chapitres contiennent des compléments mathématiques sur les thèmes 10, 14, 15 et 16 et un listing complet des codes Matlab utilisé pour les illustrations numériques

*par James Ledoux, INSA de Rennes*