« FONDAMENTAUX EN STATISTIQUE », UN MOOC DE FRANCE UNIVERSITÉ NUMÉRIQUE (FUN)

Jean-Louis PIEDNOIR[1]

TITLE

"Fundamentals in statistics", a MOOC of France Université Numérique (FUN)

RÉSUMÉ

Dans ce « libre propos », Jean-Louis Piednoir donne ses impressions sur son expérience de participation au MOOC « Fondamentaux en statistique » d'Avner Bar-Hen sur la plate-forme FUN (France Université Numérique).

Mots-clés : *MOOC, FUN.*

ABSTRACT

In this article, Jean-Louis Piednoir comments his participation to the MOOC "Fundamentals in statistics" of Avner Bar-Hen on the platform FUN (france Université Numérique).

Keywords: *MOOC, FUN.*

1 Le cadre institutionnel

La création par quelques universités états-uniennes des « massive open online courses » (MOOC) a suscité des initiatives comparables en France. En particulier, le Ministère de l'Enseignement Supérieur et de la Recherche a créé *France Université Numérique* (FUN). Parmi les cours disponibles, l'un d'entre eux, intitulé « *Fondamentaux en statistique* », a été animé par Avner Bar-Hen, professeur à l'Université Paris-Descartes. Il a duré 5 semaines et est orienté vers les applications de la statistique à la biologie.

2 L'organisation pédagogique

La durée du MOOC est de 5 semaines. Chaque semaine comprend environ 4 cours, chacun terminé par un « quiz », série d'exercices d'application immédiate du cours. Les cours sont oraux, dispensés par le professeur Avner Bar-Hen, avec apparition dans une fenêtre d'un tableau quand ce dernier est indispensable à la compréhension de l'exposé. Le même cours existe aussi sous forme de fichier PDF.

A la fin de la semaine, il est proposé un problème à résoudre et à envoyer pour correction, la solution devant tenir en une ou deux pages. Cette correction est faite en partie par d'autres inscrits au cours. Une fois le délai de correction passé, la solution est envoyée aux inscrits. Il existe aussi un forum où l'on peut poser des questions auxquelles répond l'équipe pédagogique.

[1] Inspecteur général honoraire de l'Education Nationale, amjl.piednoir@orange.fr

3 Le contenu

Le programme suivi chaque semaine est résumé ci-dessous.

1. *Statistique descriptive, cas d'une variable* : types de variables, représentations graphiques, indices de position et de dispersion, boite à moustache.
2. *Statistique descriptive, cas de variable multidimensionnelle* : tableau de contingence, corrélation, corrélation de rangs, analyse en composantes principales (ACP), analyse factorielle des correspondances (AFC).
3. *Apprentissage et classification* : recherche de groupes, classification hiérarchique ascendante, méthode des nuées dynamiques, classement des individus dans des groupes avec utilisation de la validation croisée pour mesurer la qualité de la classification.
4. *Théorie des tests* : quelques rappels de calcul des probabilités, principe d'un test, les deux hypothèses (simples ou composées), notion de niveau et de puissance, présentation de 3 tests non-paramétriques (test des signes, tests de Wilcoxon à un et à deux échantillons).
5. *Tests paramétriques* : la loi dite normale, son importance, estimation des paramètres inconnus, tests du χ^2, de Fisher, de Student.

4 Quelques remarques

L'auteur de ces lignes a été inscrit au MOOC, a suivi les cours, fait les « quiz », étudié les problèmes, mais n'a pas été jusqu'à la rédaction de leurs solutions et ne peut pas fournir d'appréciation sur la qualité des corrections par les pairs.

4.1 A propos de l'informatique

Pour l'informatique il y a souvent une période de rodage du serveur mais aussi des utilisateurs. Quelques dysfonctionnements mineurs ont été observés.

4.2 Sur la forme

Sauf réserves exposées ci-dessous, j'ai apprécié la qualité de l'exposé oral d'Avner Bar-Hen, exposé toujours très clair, agréable à écouter, avec une bonne diction sauf en quelques instants très rares. Les concepts introduits sont illustrés d'exemples, la plupart bien choisis. Les « quiz » permettent de vérifier leur compréhension, à une exception près : une erreur sur celui relatif à la loi binomiale, au départ on peut croire à un piège, ce qui n'est pas le cas ! Le vocabulaire utilisé est simple mais la présence de plusieurs anglicismes est-elle toujours indispensable ? On pense à « box-plot » pour boîte à moustaches, « consistant » pour convergent, « localisation » pour position, par exemple. Les tableaux introduits dans les cours sont clairs et parfaitement lisibles.

5 Les contenus présentés, une première analyse

Les remarques qui suivent traduisent les observations et le ressenti de l'auteur. Il est bien évident que d'autres participants peuvent avoir d'autres points de vue.

5.1 Des exemples bien choisis

Les exemples choisis pour illustrer sont de deux types, simples mais artificiels, ou extraits de données réelles. Dans les deux cas ils sont pertinents et illustrent bien les notions exposées. Il en est de même pour ceux présents dans les « quiz ».

5.2 Une progressivité des exercices de fin de semaine

A la fin de chaque semaine est proposé un devoir à rendre, qui pourra être soumis à des pairs. Les données numériques sont fournies sous la forme d'un fichier « Excel ». D'une semaine à l'autre la complexité croît, et on revient souvent sur des notions vues les semaines précédentes. Quand le délai de remise du devoir est passé, un corrigé est disponible. Ceux-ci sont clairs et en général bien illustrés de représentations graphiques.

5.3 Une mise en évidence des difficultés

Dans la partie consacrée à la statistique descriptive, on insiste à juste titre sur l'importance des représentations graphiques et les principales difficultés sont signalées : présence ou non de valeurs aberrantes, pièges engendrés par le regroupement des données en classes. Les contradictions qui existent entre les différentes qualités souhaitées pour des résumés statistiques sont bien mises en évidence.

5.4 Parfois une ambition trop élevée

C'est le cas pour le chapitre **«** *apprentissage et classification* **»**. Beaucoup de notions, de procédures sont exposées dans un temps très court, ce qui nécessite des raccourcis ne facilitant pas la compréhension. Par exemple, était-il utile à ce niveau d'introduire la validation croisée ?

Pour les *tests non-paramétriques*, le raccourci risque d'induire des contre-sens dans l'esprit du lecteur non averti, comme par exemple quand on compare la puissance de deux tests. Rappelons que le test du signe est le test sans biais le plus puissant pour tester une médiane nulle quelle que soit la forme de la distribution. Le test de Wilcoxon à un échantillon lui est préférable si la distribution sous-jacente est symétrique, ce qui n'est pas indiqué. La phrase « le test 1 est plus puissant que le test 2 car on rejette H_0 avec le test 1 et on l'accepte avec le test 2 » mériterait une clarification car, posée abruptement, elle est fausse.

5.5 Une approche intéressante des tests

La façon de présenter un test statistique comme un procès au sens juridique du terme est bien venue. Ensuite l'exposé sur les tests gaussiens, en lien avec l'estimation par intervalle de confiance, est clair et permet au néophyte de procéder à de premières applications. Une question : fallait-il mettre des rappels de calcul des probabilités dans le corps du cours ou plutôt considérer ces notions comme des prérequis ?

6 Pour le prochain MOOC

Les quelques réserves exprimées ci-dessus n'induisent pas un jugement négatif de l'auteur. Le MOOC m'est apparu de grande qualité même si certaines améliorations sont possibles. On perçoit l'importance et la qualité du travail accompli pour la mise en place des différentes composantes du MOOC. D'autres que les premiers participants devraient y avoir accès. Quand ?