

# AUTOAPPRENTISSAGE ADAPTATIF ET PERSONNALISE A L'ANALYSE STATISTIQUE EN UTILISANT R : PERSONNALYSE R

Roch Giorgi, Robert Darlin Mba, Farid Djenad, Germain Rolla, Jean-Charles Dufour

*Aix Marseille Université, INSERM, IRD, SESSTIM, Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale, Marseille, France. [roch.giorgi@univ-amu.fr](mailto:roch.giorgi@univ-amu.fr)*

**Résumé.** L'enseignement de la démarche de l'analyse statistique dans le cadre d'un enseignement à distance, de niveau master, regroupant des apprenants inscrits en formation initiale ou en formation continue, ayant différents niveaux de connaissances et de compétences est un défi pédagogique qui ne peut être relevé sans la mise en place d'une méthode et d'outils adaptés à ce public hétérogène. De plus, des objectifs de formation en termes d'acquisition de compétences concernant la mise en œuvre d'une analyse de données issues du domaine de la santé imposent : (i) d'être capable d'opter pour la bonne démarche d'analyse statistique en fonction des données auxquelles on est confronté et (ii) d'utiliser un logiciel d'analyse statistique, et donc d'en maîtriser son fonctionnement. Pour cela, nous développons une plateforme web d'auto-apprentissage adaptatif et personnalisé à la démarche de l'analyse statistique en utilisant R.

Pour mettre en place notre plateforme, nous avons développé : (i) un modèle de connaissances et de compétences composé de 4 états, eux-mêmes pouvant intégrer différents niveaux de difficultés ; (ii) un modèle de « méta-exercices » afin de s'affranchir de la création fastidieuse de multiples exercices mais en gardant toutefois la possibilité d'en générer virtuellement une infinité, différents entre eux ; et (iii) un moteur de génération d'exercices accessibles sur une plateforme web.

Les domaines de l'analyse statistique couverts à l'issue de ce projet concernent l'analyse univariée, l'analyse bivariée, l'analyse multivariée à l'aide des modèles de régression les plus classiques.

L'évaluation de ce dispositif auprès des apprenants est à prévoir afin de pouvoir quantifier leur satisfaction et leur réussite.

**Mots-clés.** Analyse statistique, plateforme d'autoapprentissage, enseignement adaptatif, enseignement personnalisé, logiciel R

**Abstract.** The statistical analysis approach training in the context of e-learning, at master's level, involving learners enrolled in initial or in continuing training, with various levels of knowledge and skills is a pedagogical challenge. It cannot be raised without introducing method and tools adapted to this heterogeneous public. Furthermore, training objectives in terms of skills training for the implementation in practice of an analysis of data from the field of health require: (i) to be able to determine the right statistical analysis approach according to the data and (ii) to use statistical software, and therefore to be able to practice. To do this, we develop a web platform of self-learning, adaptive and personalized, for the statistical analysis approach using R.

To implement our platform, we have developed: (i) a model of knowledge and skills having 4 states, themselves having the possibility to integrate various levels of difficulties; (ii) a model of "meta-exercises" in order to avoid the tedious creation of multiple exercises, but with the possibility of generating virtually an infinity, different from each other; and (iii) an engine to generate exercises accessible on a web platform.

The areas of statistical analysis covered at the end of this project concern univariate analysis, bivariate analysis, multivariate analysis using the most conventional regression models.

The assessment of this device with the learners is to be envisaged in order to be able to quantify their satisfaction and their success.

**Keywords.** Statistical analysis, self-learning platform, adaptive teaching, personalized teaching, R software.

## 1 Introduction

Enseigner la démarche de l'analyse statistique dans le cadre d'un enseignement à distance, de niveau master, regroupant des apprenants inscrits en formation initiale ou en formation continue, ayant différents niveaux de connaissances et de compétences est un défi pédagogique qui ne peut être relevé sans la mise en place d'une méthode et d'outils adaptés à ce public hétérogène. Dans nos formations, une des finalités est que l'apprenant soit capable de mettre en pratique ses connaissances et ses compétences pour effectuer une analyse de données issues du domaine de la santé. Dans ce domaine, les biostatisticiens professionnels sont tout particulièrement confrontés à des jeux de données de qualité et de natures variées (résultats biologiques, observations cliniques,...). Ainsi, il serait souhaitable que les apprenants soient également confrontés à cette variabilité afin de mettre en pratique des traitements appropriés à tel ou tel jeu de données tout en suivant une méthode d'analyse cohérente et adaptée au(x) type(s) de variable(s) à analyser (quantitative, qualitative ordonnée ou non,...). Cette mise en pratique fait bien évidemment appel à l'utilisation d'un logiciel d'analyse statistique, ce qui par conséquent ajoute un degré supplémentaire d'hétérogénéité entre les apprenants tant sur les connaissances de l'outil que sur les compétences relatives à son utilisation.

Dans le cadre du Master de Santé Publique proposé à Aix-Marseille Université, nous proposons trois spécialités complémentaires : Expertise et Ingénierie des Systèmes d'Information en Santé (EISIS), Méthodes Quantitatives et Econométriques pour la Recherche en Santé (MQERS), Santé Publique Sociétés Développement (SPSD) (<http://sesstim.univ-amu.fr>). Ces spécialités sont proposées sous un mode d'enseignement à distance et sont labellisées Académie d'Excellence par l'Idex A\*Midex (ANR-11-IDEX-0001-02) dans le cadre du projet MEDSe-PubHealth (Mediterranean-South e-Master of Public Health ; <http://sesstim.univ-amu.fr/projet/medse-pubhealth>). Ce projet a pour objectifs de proposer un programme de formation de haut niveau en Santé Publique, de contribuer à l'innovation pédagogique, d'améliorer la qualité de ces formations, de promouvoir l'offre de formation labellisée à l'international.

Pour cela, nous avons développé et étendu un modèle pédagogique de manière homogène et cohérente afin de permettre aux apprenants de suivre leur formation à distance sur un mode synchrone et avec différentes activités asynchrones. La nature interdisciplinaire de ces enseignements, la dimension internationale, l'hétérogénéité de notre public nous ont conduits à nous inscrire dans une démarche pédagogique centrée sur l'apprenant, en favorisant le développement de son auto-apprentissage, adapté à ses connaissances et à ses compétences de manière à personnaliser sa formation. C'est dans ce cadre-là que nous travaillons à la mise en place d'une plateforme web d'auto-apprentissage adaptatif et personnalisé à la démarche de l'analyse statistique en utilisant R [1]. Cette plateforme permet, à partir d'un « méta-exercice » établi par l'enseignant, de générer une multitude d'exercices différents. Les exercices ainsi générés diffèrent de par la description du contexte (l'énoncé), les libellés des questions posées et des items proposés pour répondre, mais également de par le jeu de données qui doit être traité par l'apprenant, leurs complexités et d'autres éléments permettant de moduler le niveau de difficulté.

## 2 Méthode mise en place

La réalisation de ce projet repose sur la mise en place d'un modèle de connaissances et de compétences, d'un modèle de « méta-exercices » et d'une implémentation informatique d'un « moteur » capable de générer  $n$  instances d'exercices à partir d'un seul « méta-exercice ».

- Modèle des connaissances et des compétences : il repose sur la définition d'états homogènes

de connaissances et de compétences dans le domaine de l'analyse statistique ;

- Modèle de « méta-exercices » : il comporte une partie générique, commune et réutilisable dans différents types d'analyses statistiques (univariée, bivariée, multivariée, selon le type de variables quantitatives et/ou qualitatives, ...), et une partie qui peut être spécialisée pour décrire les spécificités d'un type d'analyse en particulier ;
- Implémentation : développement d'un moteur de génération d'exercices accessibles sur une plateforme web. Il interprète la description du « méta-exercice » comme des paramètres à utiliser pour générer des instances d'exercices variables d'un apprenant à l'autre (et d'une session d'exercice à l'autre pour un même apprenant).

Les domaines de l'analyse statistique devant être couverts à l'issue de ce projet concernent l'analyse univariée, l'analyse bivariée, l'analyse multivariée à l'aide des modèles de régression les plus classiques.

### 3 Résultats

Nous avons défini quatre états homogènes (Tableau 1) :

- Etat 1 : l'apprenant est novice dans le domaine de l'analyse statistique, il n'a pas de connaissances en analyse statistique, il n'a également pas de connaissances ni de compétences dans l'utilisation de R ;
- Etat 2 : l'apprenant dispose de certaines notions dans le domaine de l'analyse statistique, il a certaines connaissances en analyse statistique, il a également certaines connaissances dans l'utilisation de R, pour lequel il n'a pas de compétences ;
- Etat 3 : l'apprenant a des bases dans le domaine de l'analyse statistique, il a des bases en analyse statistique, il a également des connaissances et certaines compétences dans l'utilisation de R ;
- Etat 4 : l'apprenant connaît le domaine de l'analyse statistique, il connaît la démarche statistique, il connaît et est compétent dans l'utilisation de R.

Il convient de préciser que l'apprenant peut connaître la démarche statistique sans toutefois maîtriser R. Il peut avoir des compétences dans l'utilisation d'autres logiciels statistiques (SAS, Stata,...).

Tableau 1 : Description des états pris en considération dans notre modèle.

Connaissances et compétences concernant	l'analyse et la démarche statistique			
		+	++	+++
l'utilisation de R		Etat 1	Etat 1	Etat 1
	+	Etat 2	Etat 2	Etat 2
	++	Non applicable	Etat 3	Etat 4
	+++			

Ces états sont déclinés selon le type de variables et selon les domaines de l'analyse statistique couverts dans le cadre de ce projet.

La manière d'appréhender l'utilisation de R dans cette démarche d'apprentissage de l'analyse statistique varie selon l'état considéré. Dans les états 1 et 2 il s'agira, à l'aide de « glisser-déposer », d'amener l'apprenant à manipuler des blocs de syntaxes R constitués d'éléments de langage déjà pré-combinés (état 1) ou des éléments de syntaxe R qu'il devra lui-même combiner (état 2). Dans les états

3 et 4 l'apprenant doit écrire lui-même la syntaxe R appropriée. Dans l'état 3 il est guidé/évalué question après question. Dans l'état 4 il est évalué sur sa capacité à écrire et utiliser dans un ordre libre mais acceptable les diverses syntaxes R utiles à la réalisation de l'analyse statistique complète et bien conduite du jeu de données auquel il est confronté.

Pour réaliser un « méta-exercice » un enseignant doit suivre le modèle que nous avons défini. Le « méta-exercice » est réutilisable pour créer différents exercices et leurs jeux de données associés. Au sein d'un « état », le niveau de difficulté d'un exercice pourra varier selon le niveau de connaissances ou de compétences d'un apprenant (par exemple selon le nombre de propositions, le nombre de tentatives autorisées, le temps imparti pour répondre, le type de distribution des variables,...). L'énoncé d'un « méta-exercice » comporte des zones pouvant prendre différentes valeurs lors de l'instanciation d'un exercice. Certaines zones sont indépendantes du jeu de données généré pour l'exercice (par exemple : synonymes d'un mot, variations grammaticales d'un syntagme,...), d'autres en revanche dépendent de ce jeu de données (par exemple : nom de la variable d'intérêt, modalités/valeurs qu'elle peut avoir,...). Toutes contribuent à la variabilité syntaxique/grammaticale des différents énoncés possibles.

Le contenu des énoncés est généré de manière automatique, avec un niveau de difficulté qui est soit choisi *a priori* par l'apprenant, soit adapté aux réponses qu'il a obtenues lors de la session en cours ou lors des sessions précédentes. Un retour cognitif est proposé par le système en cas de réponse inappropriée de sa part.

La plateforme est divisée en cinq composants applicatifs : l'application web de base (« espace apprenant »), l'application web de support (« espace concepteur »), deux services web de support (« le générateur d'instances d'exercices » et « l'adaptateur de connaissances ») et l'extension web du Serveur R « RServe ». Ces composants sont déployés sur des machines virtuelles et la communication entre eux est orientée REST. La plateforme est accessible, sur les deux applications web « espace apprenant et concepteur », via un reverse proxy.

## 4 Discussion et conclusion

Différents modèles d'enseignement à distance existent. Souvent associé à un enseignement destiné à un nombre élevé d'apprenants, voire à une masse d'apprenants, un enseignement à distance doit dispenser une formation au plus près des apprenants, de manière à ce qu'ils atteignent le même niveau de connaissances et de compétences que s'ils assistaient à un enseignement dispensé en présentiel. Au-delà de ces considérations faisant l'objet de débats, les évolutions culturelles, et professionnelles, transforment les pratiques et les attentes des apprenants vis-à-vis de leur formation. La place à l'auto-formation croît, ne serait-ce que par le développement de la pédagogie inversée. Notre modèle pédagogique a certaines similitudes avec un enseignement mixte (blended learning), mais il est entièrement à distance. Nous souhaitons renforcer la part de travail personnel des apprenants en amont des séances d'enseignement, mais aussi en aval dans le cadre d'auto-évaluation. Pour cela, nous investissons dans le développement de l'autoformation, sous un format adaptatif, personnalisé aux connaissances et aux compétences de l'apprenant, en proposant des innovations pédagogiques, notamment dans le domaine de la vidéo adaptative [2] et d'une plateforme web de génération d'exercices adaptatifs.

Le modèle cognitif et informatique que nous développons a pour objectif d'apprendre la démarche de l'analyse statistique en utilisant R. Le contenu des exercices générés par le dispositif s'adapte au niveau de connaissances et de compétences de l'apprenant. A terme, ce dispositif pourra être utilisé pour un enseignement construit selon une approche de pédagogie inversée, mais aussi suite à la dispensation de l'enseignement afin que l'apprenant puisse autoévaluer ses acquisitions. L'évaluation de ce dispositif auprès des apprenants est à prévoir afin de pouvoir quantifier leur satisfaction et leur réussite.

## **Bibliographie**

[1] R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.

[2] Dufour JC, Emeric A, Giorgi R (2016). Interactive Videos within an e-Learning Context. *Stud Health Technol Inform*, 221:117.

## **Remerciements**

Ce travail a bénéficié d'une aide du gouvernement français au titre du Programme Investissements d'Avenir, Initiative d'Excellence d'Aix-Marseille Université - A\*MIDEX.