Formation en ligne à la statistique médicale

Marcos Perreau Guimaraes[1,2], Bernard Ycart[2] et Claudine Robert[1]

[1]Projet IS2, INRIA Rhône-Alpes
[2]Equipe PRISME, Université Paris 5 René Descartes

Abstract
In the interactive teaching of statistics one has to multiply experiments on real data, as well as simulated data based on pseudo-random numbers. It is a necessary condition for a concrete understanding of the main statistical notions. This approach is illustrated by SMEL, an interactive online course of medical statistics. It is structured in four layers, connected through hypertext links: a set of articles, an index, a set of lecture notes, and a set of simulations. Many interactive experiments illustrate the main notions. The pedagogical objectives, the structure of the course and some experiments illustrating our approach are described.

Keywords

online interactive teaching, medical statistics, simulation

1 Introduction

Plus que dans d'autres domaines, l'ordinateur s'est imposé très tôt dans l'enseignement de la statistique. Cela tient sans doute à la pression des applications : depuis longtemps, plus aucun traitement de données ne se fait sans un logiciel spécialisé, et l'apprentissage de tels outils fait légitimement partie de la plupart des cursus universitaires en statistique. Ce n'est pas pour autant que le rôle de l'ordinateur dans la didactique de la discipline est pleinement compris ni formalisé.
Une expérience de 15 ans d'utilisation de la simulation dans l'enseignement des probabilités et des statistiques nous a amenés à développer une réflexion de fond sur l'apprentissage par ordinateur dans ces disciplines. Reconnaissant l'intérêt de la simulation pour l'illustration concrète d'un cours, nous avons progressivement évolué vers des outils permettant une participation plus active de la part des étudiants. Le projet SMEL (Statistique Médicale En Ligne) est en quelque sorte l'aboutissement (sans doute provisoire) de cette réflexion. Il s'agit d'un cours en ligne gratuit, et téléchargeable pour une utilisation locale, à l'adresse : `http://www.math-info.univ-paris5.fr/smel`.
Notre problématique était de fournir un outil d'auto-formation destiné à un public spécifique (médecins et étudiants en médecine). L'objectif principal est de développer une compréhension concrète et applicable des notions de base en statistique, en s'appuyant sur des expériences interactives, utilisant des données réelles ou des nombres pseudo-aléatoires.
La question de fond à laquelle SMEL apporte une réponse (subjective et partielle) est : que peuvent apporter l'utilisation de l'ordinateur et l'interactivité dans l'apprentissage de la statistique ? Il s'agit évidemment de préciser le progrès par rapport à une présentation classique. Ce progrès réside à nos yeux :

- dans la possibilité de navigation en graphe non-arborescent, grâce aux liens hypertextes,
- dans la possibilité de réaliser en peu de temps de multiples traitements de données, réelles ou simulées,
- dans l'action directe que l'on peut avoir sur les expériences par l'interface graphique.

Ces trois points seront développés et commentés dans ce qui suit. Il est clair que notre vision n'est ni nouvelle ni unique. De nombreux logiciels ou sites d'enseignement ont adopté avant nous des démarches analogues, et nous nous en sommes inspirés. En ce qui concerne plus particulièrement la possibilité d'agir directement sur les situations expérimentales par une interface visuelle, notre approche est comparable à celle du logiciel d'apprentissage de la géométrie Cabri-géomètre (voir par exemple [1,2]), qui l'a puissamment développée avec le succès mérité que l'on sait. D'autre part, notre propre expérience dans la réalisation de logiciels de simulation pour l'enseignement, et en particulier le développement d'un environnement de simulation en Mathematica pour le cours de Denker et Woyczynski [3], nous ont permis d'affiner notre vision de l'apprentissage par ordinateur.

Le choix de la mise en ligne du cours, et de l'utilisation des langages html et Java s'imposait pour des raisons de diffusion. Néanmoins les expériences interactives auraient pu être codées dans d'autres langages que Java et le cours est indépendant des propriétés du réseau internet. D'ailleurs, pour éviter les lenteurs de communication, il est proposé à l'utilisateur de télécharger l'ensemble du site, de manière à l'utiliser localement, hors connexion réseau. Le logiciel de navigation (par exemple Netscape ou Internet Explorer) joue donc dans ce cas le rôle d'une simple interface d'interprétation du texte et des activités proposées.

Dans la deuxième partie, nous décrivons les objectifs et le contenu du cours. Sa structure sera présentée dans la troisième partie. La quatrième partie sera consacrée aux expériences interactives, proposées sous forme d'applets Java.

2 Objectifs pédagogiques

Nous décrivons ici les objectifs fixés au démarrage du projet (septembre 98). Ces objectifs étaient liés au public visé, aux exigences de l'apprentissage des statistiques, et à la nécessité de développer un outil interactif et convivial.

2.1 Public et contenu

Notre but était de répondre à la demande de formation en statistique des médecins des pays francophones, en leur fournissant d'une part un outil de lecture et de compréhension de la partie statistique des articles de la littérature médicale, d'autre part un système d'aide au traitement de leurs propres données. Pour l'instant SMEL est limité à la statistique élémentaire (modèles probabilistes, statistique descriptive, estimation paramétrique et tests). Ce corpus de connaissance suffit pour acquérir le mode de pensée propre à la statistique, mais il est envisageable que le contenu évolue dans le futur pour inclure des techniques de niveau plus poussé.
Les usages de SMEL peuvent être variés. Un médecin en libéral pourra travailler aux heures qui lui conviennent quelques termes à partir du lexique ou analyser à travers les articles comment se pratique la statistique. Un étudiant en médecine pourra acquérir

rapidement une expérience numérique et une bonne connaissance des lois usuelles. Un enseignant pourra illustrer ses propos à l'aide d'exemples interactifs, et progresser dans sa pédagogie grâce aux réactions des étudiants. Nous avons conçu le cours de manière à ce que chacun puisse l'utiliser de manière partielle, en suivant ses préférences, en fonction de ses besoins et de sa progression. Aucun parcours particulier n'est mis en relief ni recommandé. Chacun pourra selon son goût utiliser librement les articles, les activités de simulation, les notes de cours ou le lexique sans qu'il y ait de passage obligé.

2.2 Pédagogie de la statistique

L'apprentissage d'une discipline comme la statistique appliquée, s'adressant à des non mathématiciens, doit beaucoup plus s'appuyer sur des exemples concrets et des simulations que sur un cours magistral. En ce qui concerne les exemples, nous avons choisi de reproduire, avec l'autorisation de leurs auteurs, des articles de la littérature médicale, illustrant l'utilisation dans de vraies applications des notions introduites. Nous avons également ajouté des articles « généralistes », qui peuvent être vus au choix comme des lectures de complément, ou des introductions de motivation. Quelle que soit la qualité d'un cours de statistique, on n'acquiert cette compétence particulière qui fait le vrai statisticien, ce « flair » qui permet de « sentir les données » qu'en manipulant et en traitant de nombreux échantillons. La simulation à l'aide d'un générateur pseudo-aléatoire est en ce sens un outil idéal car elle permet de multiplier les essais sur un modèle particulier, de créer des échantillons de taille arbitrairement grande, et de renouveler à volonté les expériences de calcul. Cependant rien ne remplace le contact avec des données réelles. Nous avons, dans les activités proposées, recherché un équilibre entre les simulations et les expériences à partir de fichiers de données réelles. Les activités illustrant dans le lexique les notions les plus importantes, sont autant que possible basées sur de « vraies données ». Ces fichiers de données sont reproduits au format texte dans SMEL. Nous avons aussi ajouté un catalogue d'adresses internet à partir desquelles l'utilisateur peut facilement accéder à d'autres fichiers de données.

2.3 Expérimentation et interactivité

L'utilisation de l'ordinateur dans l'enseignement de la statistique se heurte classiquement à deux écueils, que nous souhaitions éviter. Le premier est la démonstration passive. Dans le cadre d'un enseignement magistral, il est intéressant et utile de montrer des simulations « boîtes noires », sur lesquelles l'utilisateur n'a comme choix d'intervention que la saisie d'un ou deux paramètres. Le risque est la passivité : on n'apprend efficacement qu'en étant actif, c'est à dire en proposant et en réalisant soi-même ses propres simulations. La démonstration d'un logiciel de la part de l'enseignant demande un acte de foi de l'étudiant qui doit croire que le codage réalise bien ce qui est annoncé, sans être en mesure de le contrôler ni de le modifier. L'alternative consiste à utiliser un environnement de calcul statistique (SAS, BMDP, SPSS,…) ou mathématique (Mathematica, Maple, Matlab, Scilab,…). Ces environnements permettent la simulation d'expériences aléatoires et le traitement des données produites sous forme de sorties graphiques, sans que la difficulté de programmation soit insurmontable. L'enseignant peut alors coder « en direct », ligne par ligne, l'expérience qu'il propose, et les étudiants peuvent eux-mêmes réaliser des simulations comme travaux pratiques. Mais les compétences nécessaires à l'utilisation de ces logiciels ne sont pas partagées par tous les apprentis statisticiens. L'expérience montre que même si les étudiants ont déjà reçu une formation à leur environnement de calcul, ils passeront beaucoup plus de temps à chercher les erreurs dans leurs lignes de code, qu'à

véritablement observer et comprendre leur expérience de simulation. Un principe de base en pédagogie impose de ne pas superposer de difficulté parasite à l'objectif que l'on poursuit. Une expérience de simulation, pour être efficace, doit être immédiate à exécuter. C'est le but que nous avons cherché à atteindre en codant les expériences proposées dans SMEL de façon que leur utilisation soit la plus transparente possible. Face à une expérience, l'utilisateur doit comprendre immédiatement comment il peut agir (curseurs, zones modifiables…) et se concentrer sur l'effet de son action, qu'il doit comprendre et interpréter correctement. Nous avons toujours privilégié l'approche la plus visuelle possible : modifier une courbe en continu grâce à un curseur de contrôle des paramètres est plus facile et plus parlant que de saisir ces paramètres dans des zones de texte.

3 Structure du cours

Comme nous l'avons déjà indiqué, nous n'avons pas souhaité imposer de parcours linéaire à l'utilisateur, qui doit pouvoir naviguer à volonté dans les différentes pages, qu'il s'agisse des articles, du lexique, des notes de cours, ou des activités de simulation. Nous ne détaillerons pas ici la structuration informatique que nous avons choisie pour automatiser au maximum la génération des pages html. Cette structuration a été conçue de manière à être adaptable à n'importe quel cours en ligne, et est décrite dans [4]. Le cœur de la structure est une base de données lexicale, qui contient les mots significatifs du cours, faisant chacun l'objet de pages spécialisées, et sont reconnus comme liens hypertextes dans l'ensemble des autres pages (simulations, notes de cours et articles).

3.1 Pages de lexique

Tous les termes de statistique apparaissant dans le cours sont rassemblés dans un index qui pour l'utilisateur peut faire office de dictionnaire autant que de plaque tournante pour l'accès aux autres pages. Pour des raisons de standardisation informatique autant que de confort d'utilisation, il était souhaitable d'uniformiser autant que possible le traitement de ces termes. Il n'était cependant pas envisageable de les traiter tous de manière identique. En premier lieu parce que certains termes ont des synonymes qui ne justifient pas de page séparée. C'est ainsi que *déviation standard* et *erreur standard*, renvoient sur la même page que *écart-type*. En dehors du cas des synonymes, nous avons distingué trois types de termes :

- les *mots développés* sont chacun illustrés par une expérience interactive (la version actuelle en comporte 67). Outre cette expérience, la page d'un mot développé présente une définition brève et aussi peu mathématique que possible, et des liens vers les articles, les pages de cours où le mot apparaît, et les autres mots du lexique qui en sont proches.
- les *mots simples* ont aussi une page contenant une définition et les mêmes liens que précédemment, mais sans expérience. Par exemple les deux notions de *variance empirique* et d'*écart-type empirique* sont proches, mais seul l'écart-type, parce qu'il s'exprime dans la même unité que les données, se prête bien à la visualisation. *Variance empirique* est donc un mot simple, qui a parmi ses corrélats le mot développé *écart-type empirique*.
- les *mots nœuds* sont des mots reconnus dans les pages du cours, mais insuffisamment précis. Par exemple une *moyenne* peut être une *moyenne empirique*, une *moyenne mobile*, ou une *moyenne élaguée*. La page d'un mot nœud comme *moyenne* ne propose qu'une fenêtre déroulante, qui contient les liens vers les pages de termes plus précis.

Les pages des mots du lexique sont automatiquement engendrées par un logiciel programmé à cet effet. Ce logiciel prend en entrée la base de données lexicale, qui est un fichier texte, et engendre les pages en remplissant avec les champs de la base de données, des « patrons » faisant office de modèles de pages html pour chacun des trois types de mots (développés, simples et nœuds).

3.2 Articles

Pour le public d'étudiants en médecine et de médecins auxquels SMEL est destiné, il nous a semblé souhaitable de motiver et d'introduire les notions de statistiques par des articles directement tirés de la littérature médicale. Après autorisation des auteurs, ces articles ont été reformatés pour une présentation uniforme, puis édités de manière automatique, de sorte que les termes de statistique deviennent des liens hypertextes, renvoyant aux pages du lexique. Aux articles de médecine ont été ajoutés des articles généralistes, donnant des exemples d'applications des probabilités dans des situations simples, les plus proches possibles du bagage expérimental de chacun. En tout 26 articles figurent dans la version actuelle, mais la structure informatique est conçue de manière qu'il soit facile d'en rajouter, ce qui sera fait dans des versions ultérieures.

3.3 Notes de cours

Dans notre vision initiale du projet, nous avions pensé limiter SMEL aux articles, au lexique, et aux expériences interactives. Pour des raisons pédagogiques, autant que d'honnêteté vis à vis des utilisateurs, il nous paraissait nécessaire d'exposer les développements mathématiques qui sous-tendent chaque notion importante et qui justifient les définitions ou les théorèmes dont on utilise les applications. Il était prévu d'introduire ces développements mathématiques comme des pages annexes, rattachées à chacune des pages de termes développés. Ceci s'est avéré impraticable, pour des raisons de structuration logique et informatique. Ecrire une page de compléments pour chaque terme aurait conduit à de nombreuses répétitions, et aurait empêché qu'apparaissent clairement les fils directeurs qui relient les notions entre elles. Nous avons finalement décidé de rédiger des notes de cours, structurées comme un livre traditionnel, mais utilisant les liens hypertextes. Ce cours a été écrit en quatre parties, largement indépendantes entre elles :

- Modèles probabilistes
- Statistique descriptive
- Estimation paramétrique
- Tests statistiques.

Le contenu est classique, et se retrouve dans les très nombreuses références du domaine. A part [3] déjà cité, nous avons surtout utilisé [5] et [6]. Un parti pris très net a été adopté, en particulier dans la première partie, pour orienter le cours vers la génération de nombres aléatoires et la simulation des modèles. En revanche, les aspects les plus théoriques de la statistique mathématique, considérés comme hors de nos objectifs compte tenu du public visé, ont été laissés de côté.

Les quatre parties ont été dactylographiées en latex, puis traduites en html, en utilisant le logiciel libre latex2html [7], qui introduit automatiquement les liens hypertextes correspondant aux formules, aux définitions, théorèmes et propositions, ainsi qu'aux différents paragraphes du texte. Les pages html générées automatiquement ont ensuite été reprises par un logiciel spécialement conçu pour y introduire les liens vers les mots du lexique.

3.4 Simulations

Comme déjà indiqué, nous avons cherché à maintenir un équilibre entre l'expérimentation sur des données réelles, qui est réalisée dans les activités des mots développés, et la simulation à base de nombres pseudo-aléatoires, qui seule peut permettre de bien comprendre les comportements asymptotiques, les particularités des différents modèles, voire la notion même de modèle. Nous avons donc conçu un ensemble de 16 activités de simulation, permettant d'illustrer l'ensemble des notions du cours, des modèles probabilistes et représentations graphiques de statistique descriptive, jusqu'aux estimations et tests les plus classiques. Chacune de ces activités est présentée dans une page séparée, contenant une explication la plus concise possible. C'est à dessein que nous ne donnons pas de longue description ni d'indications détaillées sur le déroulement de l'expérience. Pour les simulations comme pour les expériences de mots développés, nous avons cherché à susciter la réflexion, à provoquer la curiosité et les interrogations. Il n'était donc pas question de livrer des activités « boîtes noires » dans lesquelles l'utilisateur assisterait passivement au déroulement du programme. Dans ce que nous proposons, des courbes apparaissent, des curseurs permettent de les faire bouger en continu, des boutons multiplient les tirages d'échantillons. L'objectif du travail pour l'utilisateur est d'agir sur les boutons et les curseurs jusqu'à comprendre les résultats observés, puis être capable de les anticiper.
Dans les pages de simulation, comme dans le cours ou les articles, les termes de statistiques sont des liens hypertextes renvoyant aux pages de lexique. Réciproquement, dans toutes les pages, des cadres de liens permettent d'accéder directement aux pages d'accueil de toutes les parties, et en particulier aux simulations.

4 Expériences interactives

Les expériences interactives, qu'elles soient basées sur des données réelles comme dans la plupart des pages de mots développés, ou sur des simulations à base de nombres pseudo-aléatoires, sont la caractéristique essentielle de notre projet. Nous espérons que c'est ce qui en fera l'originalité et l'efficacité pédagogique. C'est la raison pour laquelle nous décrivons ici leur fonctionnement en insistant sur l'interface utilisateur, de manière à préciser nos objectifs pédagogiques. Pour des raisons de standardisation du codage et de confort d'utilisation, nous avons cherché à uniformiser les interfaces, en faisant en sorte pour les mots développés que les expériences portant sur des termes liés logiquement, comme *moyenne empirique* et *écart-type empirique* par exemple, apparaissent également proches à l'utilisateur. A titre d'exemple, nous décrirons ci-dessous deux activités. La première illustre le mot développé *moyenne empirique*, la seconde est la simulation permettant d'observer les caractéristiques théoriques et empiriques des lois de probabilité classiques en dimension 1.

4.1 L'expérience « moyenne empirique »

Pour illustrer les notions de base du cours par des expériences, la première étape consiste à fixer les objectifs pédagogiques liés à la compréhension de la notion. Dans le cas de la moyenne empirique, l'objectif essentiel est d'acquérir une perception concrète de la variabilité de la moyenne, en fonction de la taille de l'échantillon, et de la distribution des valeurs de celui-ci (sensibilité aux valeurs extrêmes en particulier). D'autres objectifs pourraient être fixés, comme par exemple celui de comprendre l'associativité des moyennes. Nous avons choisi de nous limiter dans chaque expérience à un objectif

principal afin de ne pas surcharger les représentations graphiques. Pour permettre la multiplication des essais, nous avons choisi en général d'intégrer au code un fichier de données suffisamment grand pour permettre le tirage de sous-échantillons significatifs. Dans notre exemple, les données sont les tailles en centimètres de 375 enfants de 6 ans. Deux curseurs, doublés par des zones de texte modifiables, permettent de fixer les bornes, « nmin » et « nmax » d'un sous-échantillon. Ce sous-échantillon est affiché en rouge, alors que les autres données s'affichent en bleu. La moyenne empirique est calculée et affichée par un trait horizontal à l'intérieur des bornes de l'échantillon. La multiplication des essais est facilitée par un bouton « mélange » qui permute au hasard l'ensemble des données. On peut ainsi effectuer presque instantanément de multiples calculs de moyennes, pour des sous-échantillons de même taille, et observer l'évolution des résultats : des sauts importants pour les échantillons de taille faible, une quasi stabilité pour des échantillons de l'ordre de la centaine d'individus. La valeur numérique de la moyenne calculée est affichée au-dessous du graphique.

La démarche décrite ci-dessus reste valable pour un grand nombre d'activités, qu'il s'agisse comme ici de statistique descriptive ou de notions plus évoluées comme les intervalles de confiance ou les p-valeurs de tests. Nous sommes persuadés que la multiplication rapide d'expériences sur des sous-échantillons est le meilleur moyen d'acquérir une compréhension concrète des notions statistiques, et donc d'atteindre les objectifs de notre cours.

4.2 Représentations graphiques d'échantillons simulés

La compréhension de la notion de « modèle probabiliste » est un des objectifs essentiels d'un cours de statistique. Un des fondements de la statistique est en effet l'hypothèse que des données observées peuvent être vues comme des réalisations de variables aléatoires, tirées selon une certaine loi de probabilité. Les activités de simulation permettent de comprendre ce passage des lois de probabilités abstraites aux échantillons numériques. Nous présentons ici une simulation de lois unidimensionnelles, dont l'objectif est de visualiser les modèles classiques de variables aléatoires réelles. L'interface de cette expérience est divisée en quatre cadres graphiques dont les données sont communes, et qui présentent respectivement la densité (lois continues) ou le diagramme en bâtons (lois discrètes), la fonction de répartition, le diagramme en boîte, la fonction quantile.
Une fenêtre déroulante permet de choisir une famille de lois. Les quatre représentations d'une loi de cette famille s'affichent simultanément en rouge. Un ou deux curseurs dans le premier cadre permettent de modifier le ou les paramètres de la loi. Agir sur ces curseurs entraîne la modification en continu des quatre représentations. Un bouton permet de tirer un échantillon de la loi, dont les caractéristiques empiriques viennent se superposer en bleu aux caractéristiques théoriques. Un autre bouton permet d'augmenter la taille de l'échantillon pour observer graphiquement la convergence des fonctions empiriques vers les fonctions théoriques.
Nous ne décrirons pas ici le détail du codage. L'affichage des fonctions de répartitions et des fonctions quantiles, ainsi que les simulations d'échantillons font appel à des algorithmes numériques classiques. Nous avons utilisé les codes C de la librairie « dcdflib » [8], que nous avons traduits en Java.

5 Conclusion

L'apprentissage interactif proposé par SMEL peut permettre à un plus grand nombre d'acteurs des systèmes de santé d'accéder à certains aspects de la statistique. Cette approche nouvelle présente de nombreux avantages par rapport aux techniques d'enseignement classiques. L'expérimentation répétée, tant sur des échantillons de données réelles que sur des simulations à base de nombres pseudo-aléatoires permet d'acquérir une compréhension profonde et concrète des principales notions de statistique.
Un outil de formation ne doit pas induire un mode de pensée unique : c'est pourquoi SMEL laisse une réelle liberté au niveau de son utilisation. Ses usages devront être expérimentés, et les retours des utilisateurs permettront de définir les améliorations futures.

Remerciements

Le projet SMEL a pu être réalisé grâce au soutien de l'Action Concertée Incitative Télémédecine, et de l'INRIA Rhône-Alpes. Les auteurs remercient François Patte pour son aide.

Références

[1] LABORDE C, LABORDE JM, « The case of Cabri-géomètre: learning geometry in a computer based environment », *in « Integrating informative technology into education », Watson, D., Tusley, D. (Eds.), p. 95-106, Chapman Hall, London, 1995.*

[2] LABORDE JM, « Des connaissances abstraites aux réalités artificielles, le concept de micro-monde Cabri. », *in « Actes des 4e journées francophones EIAO », Nicaud JF et al. (Eds.), p. 29-40, Hermès, Paris, 1995.*

[3] DENKER M, WOYCZYNSKI W, *« Introductory statistics and random phenomena »,* Birkhäuser, Boston,1998.

[4] PERREAU GUIMARAES M, YCART B, « Structuration d'un cours en ligne : l'exemple de SMEL », *Sciences et Techniques Educatives, Vol.7, n°2, p.413-426.*

[5] DEVORE JL, *« Probability and statistics for engineering and the sciences »,* Brooks/Cole, Pacific Grove,1991.

[5] DEVORE JL, *« Probability and statistics for engineering and the sciences »,* Brooks/Cole, Pacific Grove,1991.

[6] CAPERAA P, VAN CUTSEM B, *« Méthodes et modèles en statistique non-paramétrique »,* Dunod, Paris, 1988.

[7] DRAKOS N, « Text to Hypertext conversion with LaTeX2HTML », *Baskerville, Vol. 3, No. 2, p 12-15, 1993.*

[8] BROWN B, LOVATO M, RUSSEL K, « Asymptotic power calculations: description, examples, computer code », *Statistics in Medicine, Vol. 18, n°22, p.37-51, 1999.*

Adresse de correspondance

Equipe PRISME, Université René Descartes
45 rue des Saints-Pères 75270 Paris cedex
`http://www.math-info.univ-paris5.fr/prisme`