SIMULER POUR COMPRENDRE : UN DIDACTICIEL POUR L'APPRENTISSAGE DE NOTIONS DE BASE EN STATISTIQUE INFÉRENTIELLE

Rodolphe PALM[1] et Germain ALLAGBE[2]

TITLE

Using simulations to understand: a tutorial to help understand basic concepts in statistical inference

RESUMÉ

Un didacticiel, intitulé « Probabilité et Statistique : simuler pour comprendre », est proposé aux étudiants pour les aider à comprendre diverses notions de base de la statistique. Ces notions sont présentées à l'aide de résultats de simulations, ce qui permet de les rendre plus concrètes. Le didacticiel se compose des huit modules suivants :

- notion de probabilité ;
- variable aléatoire discontinue et distribution de probabilité ;
- variable aléatoire continue et distribution de probabilité ;
- propriétés relatives aux sommes de variables ;
- distribution d'échantillonnage de la moyenne ;
- estimation d'un paramètre ;
- intervalle de confiance d'une moyenne ;
- test de conformité d'une moyenne.

Il fonctionne avec le navigateur Internet Explorer et est accessible à partir du site http://www.fsagx.ac.be/si/ en cliquant sur le lien « Didacticiel et QCM » figurant dans le plan de ce site.

Mots-clés : *didacticiel, outil pédagogique, enseignement, probabilité, statistique, simulation, inférence.*

ABSTRACT

A tutorial, named « Probabilité et Statistique : simuler pour comprendre », is offered to students to help them to understand several basic concepts in statistics. These concepts are presented using simulation results, thus making them more concrete. The tutorial consists of eight modules:

- concept of probability;
- discrete random variable and probability distribution;
- continuous random variable and probability distribution;
- properties of sums of variables;
- sampling distribution of the mean;
- estimation of a parameter;
- confidence interval of the mean;
- test for conformity of the mean.

It works with Internet Explorer and is available from the website http://www.fsagx.ac.be/si/ by clicking on the link "Tutorial and MCQ" in the site map.

Keywords: *tutorial, educational tool, teaching, probability, statistics, simulation, inference.*

[1] Université de Liège (ULg) et Gembloux Agro-bio Tech (GxABT), Unité de Statistique, Informatique et Mathématique appliquées, Rodolphe.Palm@ulg.ac.be

[2] Idem, allmag54@gmail.com

1 Introduction

Notre expérience d'enseignant montre qu'à l'issue d'un premier cours de statistique fondamentale destiné à des bacheliers bioingénieurs[3] et dont le volume horaire est pourtant de 60 heures (5 crédits), beaucoup d'étudiants ne maîtrisent pas des concepts de base comme, par exemple, l'interprétation d'un intervalle de confiance ou du résultat d'un test d'hypothèse. Devant ce constat, il nous a paru utile de mettre à leur disposition un outil d'auto-apprentissage qui leur permette de se familiariser davantage avec ces concepts.

Assez naturellement le choix s'est porté sur un outil utilisant les nouvelles technologies de l'information et de la communication, afin de bénéficier, d'une part, des avantages reconnus à cette approche pour n'importe quelle matière enseignée (large diffusion, coût d'utilisation réduit, interactivité, etc.) et, d'autre part, de la puissance de calcul nécessaire à la réalisation de calculs répétitifs. Ce dernier point est, en effet, fondamental, dans la mesure où nous souhaitions proposer un didacticiel dont l'originalité est de faire un large usage de la simulation statistique.

Après cette introduction (paragraphe 1), nous examinons d'abord la conception pédagogique du didacticiel (paragraphe 2), ensuite nous en présentons succinctement les différents modules (paragraphe 3) et nous donnons quelques informations relatives à sa conception technique (paragraphe 4) avant de conclure (paragraphe 5).

Le didacticiel intitulé « Probabilité et Statistique : simuler pour comprendre » (Allagbe et Palm, 2009) peut être utilisé sans restriction et est accessible à l'adresse http://www.fsagx.ac.be/si/ en cliquant sur le lien « Didacticiel et QCM » figurant dans le plan de ce site.

2 Conception pédagogique

2.1 Approche par simulation

L'approche pédagogique utilisée dans le didacticiel repose sur la simulation. Avec l'ordinateur, il est en effet très facile de simuler un grand nombre de répétitions d'expériences, telles que le jet d'un dé, le prélèvement successif de pièces dans une production ou le tirage d'échantillons dans des populations. Les résultats de ces simulations permettent ainsi de rendre plus concrètes diverses notions fondamentales de la statistique.

Cette approche devrait faire naître chez les utilisateurs un mode de réflexion applicable à d'autres situations que celles envisagées dans le didacticiel, par simple transposition des problèmes. Ainsi, par exemple, une bonne compréhension du test de conformité d'une moyenne (test portant sur une moyenne), pris comme exemple dans le didacticiel, devrait permettre une compréhension aisée d'autres tests d'hypothèses, la démarche théorique lors de la réalisation d'un test étant fondamentalement très proche, quel que soit le test considéré. On retrouve, par exemple, de manière systématique les notions d'hypothèse nulle, d'hypothèse alternative, de risques d'erreur, de puissance du test.

[3] En Belgique, le titre de bachelier est un grade académique sanctionnant des études supérieures de premier cycle de 180 crédits.

Le didacticiel n'a donc pas comme objectif de couvrir de manière exhaustive l'ensemble des méthodes statistiques le plus souvent abordées dans les enseignements de base mais bien d'illustrer, par des exemples concrets, la démarche statistique. Dans sa version actuelle, il comporte huit modules qui sont brièvement décrits au paragraphe 3.

La plupart de ces modules sont constitués de 15 à 30 pages successives. A raison d'une minute par page environ, la durée d'exécution d'un module est donc de l'ordre de la demi-heure. Cette durée pourra cependant fluctuer très fortement en fonction du temps de réflexion de l'utilisateur ainsi que du nombre de répétitions et de retours en arrière qu'il demandera.

2.2 Structure générale des modules

Les modules commencent systématiquement par une courte présentation des objectifs poursuivis et des prérequis nécessaires à sa compréhension. Ces prérequis peuvent renvoyer à des notions examinées dans un module précédent, directement accessible par un lien, ou bien concernent des notions de base qui ne sont pas présentées dans le didacticiel. Il s'agit essentiellement d'éléments de statistique descriptive (fréquences, représentations graphiques, paramètres) et de quelques distributions théoriques (distributions normales et distributions de Student).

Chaque module fait également l'objet d'une synthèse destinée à faire le point sur les notions illustrées dans le module.

Enfin, les modules concernant l'inférence statistique se clôturent par une page permettant la réalisation de simulations complémentaires. En particulier, les modules consacrés à la distribution d'échantillonnage, à l'intervalle de confiance et au test de conformité de la moyenne proposent des simulations pour sept distributions théoriques présentant des degrés de symétrie et d'aplatissement différents. Pour ces distributions, l'utilisateur peut fixer la moyenne et l'écart-type de la population, ainsi que la taille de l'échantillon et le nombre d'échantillons prélevés au cours de la simulation. Ces simulations complémentaires permettent notamment à l'utilisateur d'apprécier l'importance de la normalité de la population-parent comme condition d'application de plusieurs méthodes statistiques.

3 Modules proposés

3.1 Notion de probabilité

La démonstration illustre le phénomène de stabilisation (convergence) des fréquences relatives, appelées aussi simplement fréquences, qui permet de définir la probabilité comme une forme idéalisée de la fréquence relative. La probabilité de production d'une pièce non conforme par une machine est prise comme exemple.

3.2 Variable aléatoire discontinue et distribution de probabilité

La démonstration introduit les notions de variable aléatoire discontinue (discrète), de distribution de probabilité et de fonction de répartition. Ces notions sont présentées comme des formes limites des distributions de fréquences non groupées, utilisées en statistique descriptive. La moyenne et la variance d'une variable aléatoire discontinue sont également définies. Ces notions sont illustrées par l'exemple de la distribution du nombre de pièces non

conformes présentes dans un groupe de 10 pièces prélevées au hasard dans la production d'une machine.

3.3 Variable aléatoire continue et distribution de probabilité

La fonction de densité de probabilité et la fonction de répartition sont présentées comme des formes limites d'histogrammes normés et de polygones de fréquences cumulées[4]. L'accent est mis sur les relations entre les deux fonctions et sur la représentation des probabilités sur les graphiques de ces fonctions.

La longueur de pièces produites par une machine sert de support à ce module. Quatre exemples de distributions théoriques continues sont ensuite présentés dans le but de mettre en évidence les principales caractéristiques des fonctions de répartition et de densité de probabilité.

3.4 Propriétés relatives aux sommes de variables

La démonstration illustre quelques propriétés relatives aux sommes de variables aléatoires indépendantes et plus particulièrement les propriétés relatives aux moyennes et aux variances des sommes. La propriété d'additivité propre à certaines familles de variables et la convergence de la somme de variables aléatoires indépendantes vers les distributions normales sont également abordées.

Quatre exemples sont traités successivement :

- somme des résultats de deux dés ;
- somme des nombres d'incidents sur deux chaînes de production indépendantes (somme de deux distributions de Poisson) ;
- sommes des résultats de 10, 20 et 30 jets de dés ;
- sommes de 10, 20 et 30 variables très dissymétriques (exponentielles négatives de moyenne unitaire).

3.5 Distribution d'échantillonnage de la moyenne

La distribution d'échantillonnage de la moyenne est présentée comme étant la forme limite de l'histogramme normé des moyennes observées lorsqu'on répète le prélèvement, dans une population donnée, d'un échantillon de taille fixée et qu'on en calcule la moyenne. Les caractéristiques de cette distribution sont décrites et les résultats théoriques sont vérifiés empiriquement par simulation.

La population prise en compte est la population des tailles des exploitations agricoles de la Région wallonne (Belgique). Cette population a été simulée à partir de la distribution de fréquences des 17 109 exploitations provenant du recensement agricole de 2005. Sept distributions théoriques, présentant des degrés de symétrie et d'aplatissement différents, sont également proposées dans un module de simulations complémentaires.

[4] L'histogramme normé est un graphique associant à chaque classe un rectangle dont la hauteur est égale à la fréquence unitaire. Le polygone de fréquences cumulées est aussi appelé courbe cumulative des fréquences.

3.6 Estimation d'un paramètre

La méthode des moments et la méthode du maximum de vraisemblance sont présentées à partir de deux exemples : l'estimation de la borne supérieure du domaine de variation d'une distribution uniforme et l'estimation du paramètre d'une distribution exponentielle négative. Les notions de biais et de précision des estimateurs sont illustrées à partir des distributions d'échantillonnage empiriques, obtenues par simulation.

3.7 Intervalle de confiance d'une moyenne

Le principe de construction de l'intervalle de confiance de la moyenne est expliqué, l'accent étant mis sur l'interprétation de cet intervalle, en relation avec le degré ou niveau de confiance. Le cas où la variance de la population doit être estimée est également présenté.

La population prise en compte est la population des tailles des exploitations agricoles déjà considérée dans le module consacré à la distribution d'échantillonnage de la moyenne. Des simulations complémentaires peuvent être réalisées à partir de sept distributions théoriques différentes.

3.8 Test de conformité d'une moyenne

La démonstration explique le principe de construction du test, en insistant sur l'interprétation du risque de première espèce et de la fonction de puissance. L'incidence de l'effectif de l'échantillon sur le test est également examinée.

Le cas d'une machine produisant des pièces dont on mesure la longueur sert d'illustration. Des simulations complémentaires permettent, ici aussi, de mettre en évidence l'importance des conditions d'application.

4 Conception technique

4.1 Organisation générale des écrans

Les écrans ont été conçus de manière à ce que l'utilisateur sache toujours dans quelle partie du didacticiel il se trouve, d'où il vient et où il peut aller.

La figure 1 présente l'organisation générale des écrans. Les zones de titre et de sous-titre indiquent en permanence le nom du module (en blanc sur fond bleu) et du paragraphe (en rouge sur fond bleu) en cours d'étude. Une icône propre à chaque module (en haut à gauche) illustre le titre.

Un clic sur l'onglet « Plan » ouvre une fenêtre affichant le plan du module en cours, ce qui permet à l'utilisateur de repérer facilement la structure du module et, par un clic sur un des éléments du plan, d'accéder directement au paragraphe en question.

En dessous de la zone de lecture et de travail (en noir sur fond blanc), la zone de navigation comporte différents boutons verts permettant d'accéder à la page suivante, à la page précédente, en haut de la page courante et au sommaire du didacticiel. Ce sommaire donne la liste des modules et le déplacement de la souris sur le texte d'un module fait

apparaître le plan du module et permet, par un simple clic, d'accéder directement au paragraphe sélectionné.

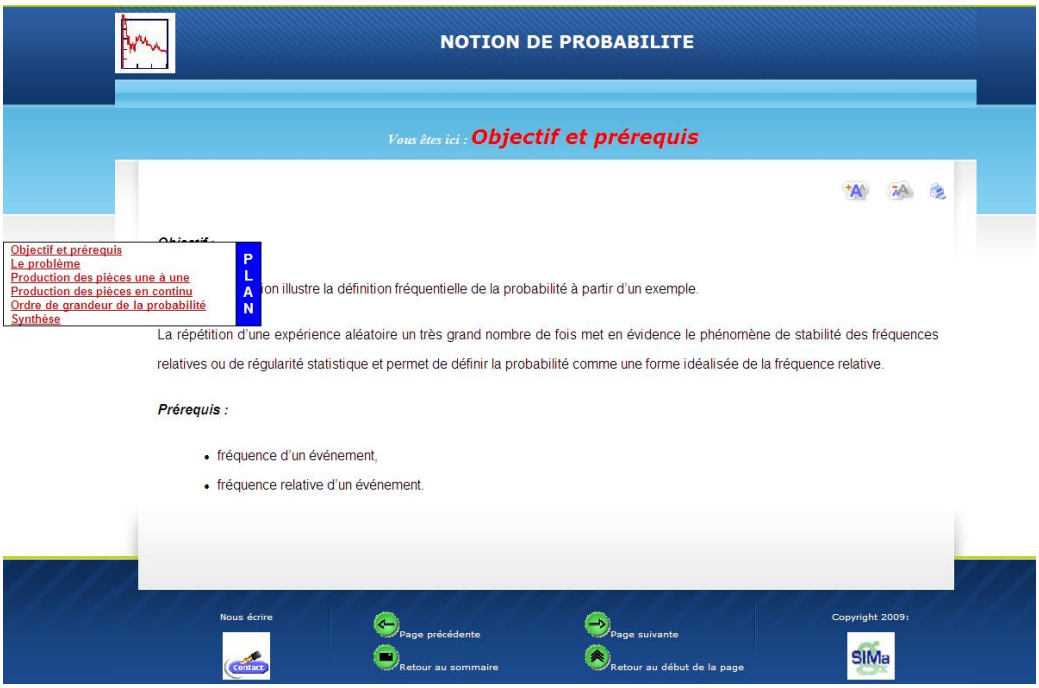

FIGURE 1 – *Organisation générale des écrans*

4.2 Interactivité

Outre la navigation proprement dite, les différentes interactions utilisées sont le déclenchement d'un traitement, la saisie de paramètres et le choix d'options. Le déclenchement d'un traitement se fait par des boutons de commande jaunes, la saisie de paramètres se fait sous la forme de zones de texte et le choix des options par des boutons d'option ou par des menus déroulants.

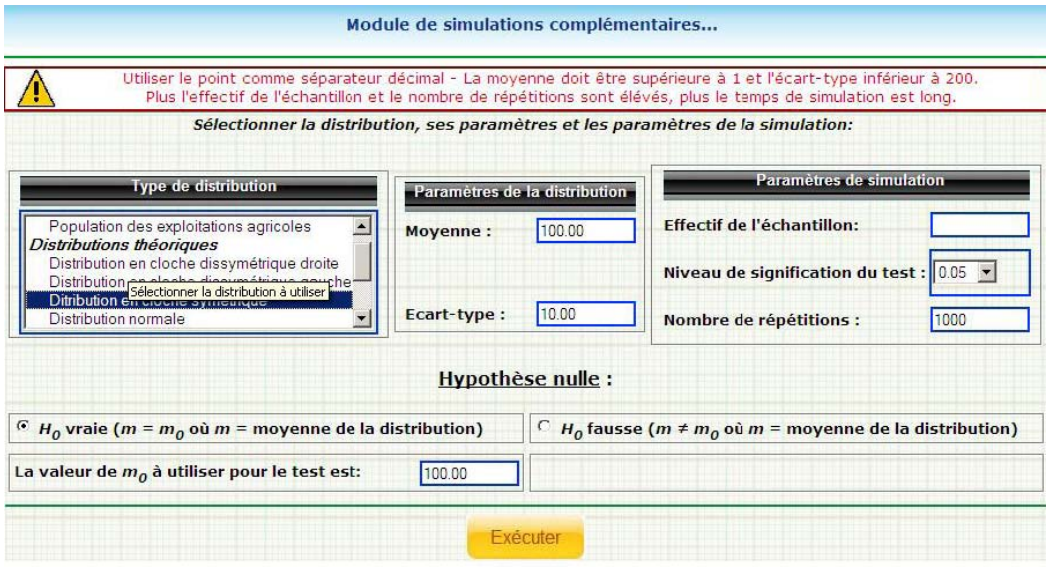

FIGURE 2 – *Eléments d'interactivité*

Pour les zones de texte, des valeurs par défaut sont le plus souvent proposées et des messages d'erreur apparaissent en cas de saisies erronées.

La figure 2 reprend une partie d'écran regroupant ces différents éléments.

4.3 Les outils et supports de développement

Le didacticiel a été développé à l'aide de plusieurs langages de programmation. Perl (Practical Extraction and Report Language) a été utilisé pour les calculs scientifiques et la génération des pages Web dynamiques. En outre, Javascript, les feuilles de style en cascade et le langage HTML (HyperText Markup Language) ont été utilisés, ainsi que le module CGI (Common Gateway Interface) comme interface du serveur Web.

En l'état actuel, l'application peut fonctionner correctement sur le navigateur Internet Explorer 7.0 avec Javascript activé.

5 Conclusions

Le didacticiel propose huit modules illustrant différents concepts de base de la statistique inférentielle, en insistant sur l'interprétation des concepts plutôt que sur le formalisme mathématique. Son utilisation ne nécessite qu'un minimum de prérequis, ceux-ci se limitant principalement aux notions de base classiques de la statistique descriptive.

Bien qu'il ait d'abord été développé à l'intention d'étudiants suivant un premier cours de statistique, ce didacticiel devrait également permettre à des utilisateurs occasionnels des outils statistiques de rafraîchir leurs connaissances en vue d'une meilleure compréhension de leurs résultats d'analyse.

Le didacticiel se focalise sur un nombre limité de concepts statistiques. Il doit être considéré comme un outil pédagogique complémentaire aux outils habituels (cours en auditoire, travaux pratiques ou dirigés, notes de cours, livres, etc.). Comme l'indique le titre du didacticiel – simuler pour comprendre –, l'approche pédagogique utilisée est très spécifique. Cette approche nous paraît particulièrement bien indiquée pour illustrer certains concepts mais elle est sans doute beaucoup moins utile pour aider à la compréhension d'autres notions. Une utilisation exagérée de la simulation pourrait d'ailleurs être contre-productive, en donnant à l'étudiant l'impression que la statistique serait avant tout une affaire de simulation et qu'il ne serait, par exemple, pas possible de réaliser une inférence statistique sans répéter un grand nombre de fois l'échantillonnage ! Pour cette raison, nous ne pensons pas qu'il soit opportun d'étendre le champ couvert par le didacticiel en ajoutant de nombreux modules nouveaux, qui seraient, comme les modules actuels, basés sur la simulation.

Enfin, d'un point de vue technique, des améliorations devraient être apportées au didacticiel afin de permettre son fonctionnement sur différents navigateurs et sur différents systèmes d'exploitation. En effet, dans sa version actuelle, ce didacticiel ne fonctionne de manière correcte qu'avec le navigateur Internet Explorer. Pour d'autres navigateurs, des imperfections dans l'affichage peuvent se produire.

Référence

[1] Allagbe, G. et R. Palm (2009), Probabilité et statistique : simuler pour comprendre (Version 1.0), Didacticiel, Université de Liège, Gembloux Agro-Bio Tech, Unité de Statistique, Informatique et Mathématique appliquées, *http://www.fsagx.ac.be/si/Didacticiel/simaccueil.pl* .