Cours Statistique et logiciel R (2015-2016)

Objectif
Au cours des dernières décennies, de nouvelles approches stochastiques ont émergé permettant d'analyser de gros volumes de données et/ou de résoudre des problèmes complexes. Les formations en statistique dispensées auprès des scientifiques sont parfois insuffisantes pour leur permettre de comprendre et de maîtriser ces nouvelles approches.

L'objectif de cette formation est de vous permettre d'acquérir le vocabulaire et de comprendre la démarche statistique, afin d'être capable d'identifier les méthodes statistiques adaptées à vos applications. Son contenu pourra être un prérequis à des formations ultérieures sur les nouvelles approches stochastiques.

La formation est proposée sous une double étiquette Collège doctoral/MaiMoSiNE (Maison de la Modélisation et de la Simulation) avec une priorité d'accès aux étudiants du collège doctoral de Grenoble.

Contenu de la formation
L'accroissement des données disponibles dans toutes les disciplines (sciences expérimentales, sciences humaines et sociales, données issues de calculs numériques) rend parfois complexe l'analyse des phénomènes étudiés. Il est alors essentiel de pouvoir extraire de l'information de ces bases de données, à l'aide d'outils statistiques.
Ce module est destiné aux scientifiques souhaitant s'approprier les bases de la statistique descriptive et inférentielle. Il se veut une initiation à la résolution de problème par une approche stochastique. Les concepts d'analyse des données, d'estimation et tests, d'analyse de la variance et de régression seront abordés d'un point de vue pratique. L'application de ces méthodes se fera à l'aide du logiciel gratuit R (http://cran.r-project.org/) et de son environnement RStudio (http://www.rstudio.com/).

L'enseignement comprend des exposés théoriques d'une à deux heures suivis de travaux pratiques basés sur les notions présentées.

Afin de mettre en pratique les notions présentées, les participants seront invités à présenter le type de données et de questions auxquels ils sont confrontés dans leur application.

Pré-requis
Connaissance de la statistique élémentaire : moyenne, fréquence, variance, écart-type, médiane, histogramme, etc

Planning (10 séances de 3h)
Après avoir introduit le logiciel R et l'environnement de travail RStudio, chaque séance sera constituée d'un cours et d'un TP permettant d'illustrer les différentes techniques statistiques.

1. Séance 1 et 2 : 22/03/2016 et 24/03/2016 de 9h à 12h (1h30 Cours ; 1h30 TP) R. Drouilhet
    - Prise en main du logiciel RStudio
    - Espace de travail, sauvegarde, aides
    - Structures en R (**vector** et **matrix**, **list** et **data.frame**, ...)
    - Représentations graphiques
    - Eléments de programmation (instructions de base et introduction à la notion d'objet)
    - Introduction aux possibilités d'extensions du logiciel R (API C, Rcpp, ...)
    - Introduction aux outils de reporting automatique (Sweave, knitr, ...)
2. Séance 3 et 4 : 29/03/2016 et 1/04/2016 de 9h à 12h (1h30 Cours ; 1h30 TP) F. Letué
    - Introduction générale à la statistique, prise en main d'un jeu de données (type de fichiers, type de variables ...)
      TP : présentation de jeux de données des participants, définition des populations, recodage de variables, etc ...

- Statistiques descriptives : quels tableaux, quels graphiques pour quel type de variables ?

3. Séance 5 : 4/04/2016 de 14h à 17h (1h30 Cours ; 1h30 TP) L. Viry
   - Analyse de données (descriptif, multivarié) : analyse en composantes principales, analyse des correspondances
4. Séance 6 : 6/04/2016 de 9h à 12h (1h30 Cours ; 1h30 TP) A. Leclercq-Samson
   - rappels de probabilités
   - statistique inférentielle
   - TP : simulation de variables aléatoires
5. Séance 7 : 11/04/2016 de 9h à 12h (1h30 Cours ; 1h30 TP) A. Leclercq-Samson
   - estimation et tests, problèmes à un échantillon
   - tests d'adéquation (notion de test) à une valeur ou à une loi
   - intervalle de confiance
6. Séance 8 : 13/04/2016 de 9h à 12h (1h30 Cours ; 1h30 TP) F. Letué
   - Problèmes à deux échantillons :
     - 2 variables qualitatives : test d'indépendance
     - une variable quantitative, une variable qualitative : comparaison de moyenne, puis ANOVA (à un facteur)
7. Séance 9 : 25/04/2016 de 9h à 12h (1h30 Cours ; 1h30 TP) A. Leclercq-Samson
   - deux variables quantitatives : régression simple et série temporelle
8. Séance 10 : 26/04/2016 de 9h à 12h (1h30 Cours ; 1h30 TP) A. Leclercq-Samson
   - Régression multiple + extensions : données manquantes, GLM, modèles mixtes, données spatiales (il s'agit ici plus d'identifier des types de données ou de problèmes que de traiter tous les sujets ...)