MASTER MATHS - Statistique Science des Données (SSD)

Semestre 7 – Statistique Science des Données – SSD			
5 ECTS	HAX709X	Processus Stochastiques	21hCM + 21hTD
5 ECTS	HAX710X	Statistique Inférentielle	21hCM + 21hTD
5 ECTS	HAX711X	Analyse des Données Multidimensionnelles	21hCM + 21hTD
5 ECTS	HAX706X	Optimisation	21hCM + 21hTD
4 ECTS	HAX712X	Développement Logiciel	12hCM + 18hTD
2 ECTS	HAX713X	Théorie de l'Information et de la Décision	9hCM + 9hTD
4 ECTS	HAI725I	Système d'Information et Base de Données	12hCM + 24hTD

Semestre 8 – Statistique Science des Données – SSD			
5 ECTS	HAX814X	Modèle Linéaire	21h CM + 21h TD
4 ECTS	HAX818X	Séries Temporelles	15hCM + 15h TD
4 ECTS	HAX809X	Estimation et Tests non paramétriques	15h CM + 15h TD
2 ECTS	HAX806X	Contrôle	9h CM + 9h TD
2 ECTS	HAX815X	Programmation R	6h CM + 12h TD
2 ECTS	HAX805L	Anglais	15h TD
5 ECTS	HAX817X	Projet	

6 ECTS à choisir parmi							
2 ECTS 2 ECTS 2 ECTS	HAR803B HAU804I HAU802I	Outils épidémio. Information bio. BILL	9hCM +9hTD	4 ECTS	HAU801I	Alignement et Phylogénie	12hCM + 24hTD
2 ECTS	HAX813X	Micro-économie	9hCM +9h TD	4 ECTS	HAX807X	Économie Générale	15hCM +15hTD

Semestre 9 – Statistique Science des Données – SSD							
SSD – BIOSTAT					SSD – MIND		
5 ECTS	HAX904X	Analyse Multivariée				21hCM	
5 ECTS	HAX907X	Apprentissage Statistique				21hCM	
5 ECTS	HAX912X	Modèles Linéaires Généralisés				21hCM	
3 ECTS	HAX916X	Projet ou Alternance					
2 ECTS	HAX906L	Anglais				15hTD	
5 ECTS	HAX908X	Estim. Non-paramétrique	21hCM	10 ECTS		Management des Risques (IAE)	84h TD
5 ECTS	HAX918X	Statistique Bayésienne	21hCM				

Semestre 10 – Statistique Science des Données – SSD							
SSD – BIOSTAT					SSD – MIND		
4 ECTS	HAX002X	Analyse des durées de Vie				18h CM	
4 ECTS	HAX006X	Modèles à variables latentes				18h CM	
4 ECTS	HAX004X	Complément 1	18hCM	4 ECTS	Stratégie d'entreprise et Gestion de Projet (IAE)	18hCM +15hTD	
4 ECTS	HAX005X	Complément 2	18hCM	4 ECTS	Data marketing Clients et Données manquantes	18hCM +15hTD	
14 ECTS	STAGE ou ALTERNANCE						

UE de la première année de MASTER SSD

UE HAX709X – Processus stochastiques

Description :

La première partie de ce cours concerne des compléments de théorie des probabilités : espérance conditionnelle, vecteurs gaussiens. La deuxième partie présente une des principales familles de processus stochastiques en temps discret les chaînes de Markov. Il s'agit de suites de variables aléatoires dépendantes, dont la relation de dépendance est relativement simple puisque chaque variable ne dépend que de la précédente. Il s'agit également d'un outil de modélisation très puissant. On étudiera les principales propriétés de ces processus, ainsi que leur comportement en temps long et l'estimation de leurs paramètres.

Objectifs :

Les objectifs du cours sont
- être capable de faire des calculs d'espérance et de lois conditionnelles
- être capable de modéliser une expérience par une chaîne de Markov
- être capable de calculer les grandeurs d'intérêt (probabilité et temps d'atteinte de certains événements)
- être capable de déterminer le comportement asymptotique du processus.

Pré-requis nécessaires : Cours de probabilité de niveau L3 : variables et vecteurs aléatoires, modes de convergence des suites de variables aléatoires, convergence des suites variables aléatoires indépendantes et identiquement distribuées. (+ fonction caractéristique si vecteurs gaussiens)

Algèbre linéaire : calcul matriciel, éléments propres, résolution de systèmes linéaires, suites récurrentes linéaires

Pré-requis recommandés : Théorie de la mesure

Programme :

1. Mesurabilité
    - 1.1 Tribus
    - 1.2 Processus aléatoires
    - 1.3 Filtration
    - 1.4 Temps d'arrêt

2. Espérance conditionnelle

2.1 Probabilité conditionnelle par rapport à un événement
2.2 Espérance conditionnelle par rapport à une tribu .
2.3 Espérance conditionnelle et indépendance .
2.4 Lois conditionnelles

3. Chaînes de Markov

3.1 Matrices stochastiques
- Définition et représentation graphique
- Classes communicantes
- Périodicité

3.2 Processus de Markov
- Définition d'une chaîne de Markov
- Propriété de Markov

3.3 Problèmes de passage
3.4 Classification des chaînes de Markov
- Récurrence et transience
- Lien avec la structure de classes

3.5 Comportement asymptotique
- Loi invariante
- Convergence vers la loi invariante
- Théorème ergodique
- Statistique des chaînes de Markov

UE HAX710X – Statistique inférentielle

Description : L'importance de la science statistique dans le processus de découverte scientifique et d'avancée industrielle est qu'elle permet la formulation d'inférences concernant des phénomènes d'intérêt auxquels on peut associer des risques d'erreur ou des degrés de confiance. Le calcul de ces risques d'erreur s'appuie sur la théorie des probabilités, mais les principes et des méthodes permettant d'associer ces risques aux inférences constituent un corpus théorique qui sert de base à l'ensemble des méthodologies statistiques.

Ce module se veut une présentation assez complète de ces principes de base et des outils, résultats et théorèmes mathématiques utilisés en statistique inférentielle. On y développe les notions d'estimation ponctuelle et par intervalle, de tests d'hypothèses et des concepts fondamentaux comme les familles exponentielles et le principe du maximum de vraisemblance et l'usage de la p-value. Pour la mise en œuvre de certaines applications, on présentera les outils adaptés du logiciel R.

Objectifs : Au terme de ce module, l'étudiant devra savoir développer les méthodologies statistiques optimales pour l'estimation et les tests d'hypothèses dans certaines familles de lois de probabilités paramétriques. Il devra comprendre les limites des inférences produites et être en mesure de les restituer auprès d'utilisateurs. Face à de petits jeux de données, il devra savoir choisir de façon raisonnée la meilleure approche et effectuer par le logiciel R les calculs nécessaires.

Pré-requis nécessaires : Un cours de calcul des probabilités de niveau Licence.

Pré-requis recommandés : Un cours de statistiques descriptives de niveau Licence serait un atout.

Programme :

1. Modèle statistique paramétrique

    a) Modèle statistique paramétrique; b) Modèle d'échantillonnage iid ; c) Rappels sur les théorèmes asymptotiques (LGN, TCL, Delta-méthode) ; d) Notion de statistique - caractéristiques empiriques d'un échantillon & lois asymptotiques.

2. Famille exponentielle

    a) Définition ; b) Moments.

3. Score et information de Fisher

    a) Score ; b) Information de Fisher ; c) Cas de la famille exponentielle.

4. Statistiques exhaustives

    a) Exhaustivité & caractérisations ; b) Statistique exhaustive minimale ; Statistique complète.

5. Estimation ponctuelle

    a) Risque. Risque quadratique = biais2 + variance. Ordre sur les estimateurs. Absence d'estimateur optimal ; b) Estimateur sans biais: inégalité de Fréchet. Estimation efficace & famille exponentielle. Amélioration de Rao. ESB optimal & théorème de Lehmann-Scheffe ; c) Estimation du maximum de vraisemblance, propriétés asymptotiques ; d) Estimation par la méthode des moments, propriétés asymptotiques.

6. Estimation ensembliste

    a) Région de confiance ; b) Pivot ; c) Région de confiance asymptotique.

7. Tests d'hypothèses

a) Problème de test: hypothèses, erreurs, pertes, risques associés, niveau et puissance. Fonction de test. Test pur vs test mixte ; b) Absence de test optimal. Test sans biais. Test convergent. c) Principe de Neyman ; d) Test d'hypothèses simples: test PP de Neyman ; e) Tests d'hypothèses unilatérales: famille à rapport de vraisemblances monotone & tests UPP ; f) Tests d'hypothèses bilatérales: famille exponentielle & tests UPPSB ; g) Lien entre régions d'acceptation d'un test et régions de confiance ; h) Tests asymptotiques: test de Wald, test des scores de Rao, test du rapport des maxima de vraisemblance.

8. Problèmes à deux échantillons : Comparaison de paramètres: estimation et tests.

9. Tests d'adéquation

a) Test du chi2 & application au test d'indépendance. b) Test de Kolmogorov-Smirnov. c) Tests de normalité de Shapiro-Wilks.

UE HAX711X – Analyse des données multi-dimensionnelles

Description : Les données statistiques ne cessent de devenir plus massives. Préalablement à leur modélisation, il est indispensable de les explorer et d'en réduire la dimension en perdant le moins d'information possible. Tel est l'objectif de ce cours de statistique exploratoire multidimensionnelle. Sur le plan méthodologique, les outils qu'il utilise sont essentiellement ceux de la géométrie euclidienne. Les problèmes et notions statistiques y seront donc traduits dans le langage de la géométrie euclidienne avant d'être traités dans ce cadre. Les deux familles de méthodes exploratoires qui seront vues dans ce cours sont: 1) les méthodes de classification automatique, qui regroupent les observations en classes et réduisent leur disparité des observations à la disparité entre ces classes; 2) les méthodes d'analyse en composantes, qui recherchent les directions principales de disparité entre les observations et permettent de fournir de cette disparité des images interprétables en dimension réduite.

Programme :

I - Introduction:

a) Données multidimensionnelles, observations, variables, codages ; b) Traductions en nuages de points en espaces métriques euclidiens. c) Nécessité d'une réduction dimensionnelle: composantes / classes.

II - Écritures géométriques de quantités statistiques

1. Description univariée:

a) Moyenne, fréquence, b) variance et écart-type. c) Centrage et réduction d'une variable.

2. Liaisons bivariées:

a) Liaison bivariée & conditionnement. b) Covariance et corrélation de deux variables quantitatives. c) R2 d'analyse de variance d'une variable quantitative sur une variable qualitative. d) Phi2 et T2 de deux variables qualitatives. e) Écriture unifiée des liaisons. f) Limites du bivarié & comment le dépasser.

III - Classification automatique

1. Dissemblance et ressemblance.

a) Mesures. b) Ressemblance partielle vs globale.

2. Ressemblance partielle: classification logique/conceptuelle par treillis de Galois.

3. Ressemblance globale:

a) Partitionnement en espace métrique: méthode des K-means & raffinements ; b) Classification hiérarchique: indices, algorithme de CAH, critères de choix de partitions ; c) Classification mixte ; d) Interprétation des classes ; e) La classification sur variables.

IV - Analyses en composantes principales

1. ACP normée

a) Nuage des individus, inertie et ACP directes. b) Nuage des variables, inertie et ACP duale. c) Relations de dualité et interprétation jointe des graphiques. d) Éléments supplémentaires & relation de dualité. e) La première composante comme estimation d'une variable latente continue.

2. ACP générale (avec métriques quelconques)

a) Nuage des lignes et ACP directe. b) Application au multidimensional scaling. c) Quelle ACP des colonnes, pour quelles relations de dualité? d) Les aides à l'interprétation. e) Éléments supplémentaires & relation de dualité. f) Formule de reconstitution (décomposition en éléments singuliers).

3. Analyse des correspondances binaires

a) Le Phi2 comme inerties directe et duale des nuages de profils-lignes et de profils-colonnes ; b) Quelles métriques pour quelles relations de dualité: les positionnements barycentriques ; c) Interprétation jointe des graphiques. d) Effet Guttman. e) Éléments supplémentaires.

4. Analyse des correspondances multiples.

a) Application de l'ACB à un tableau logique disjonctif complet ; b) Application de l'ACB à un tableau de Burt; équivalence ; c) Relations barycentriques entre individus et modalités. Relations barycentriques entre modalités ; d) Effet Guttman. e) Éléments supplémentaires. f) La première composante comme estimation d'une variable latente continue.

5. La pratique de l'ADM

a) Complémentarité de l'AF et de la CA ; b) Comment mener une bonne ADM.

UE HAX706X – Optimisation

Description : Ce cours est la continuation du cours d'optimisation du second semestre de L3. Après un rappel des résultats et méthodes numérique pour les problèmes d'optimisation d'ordre un et deux, sans contrainte et sous contraintes d'égalité et d'inégalité, le cours s'intéresse aux questions aujourd'hui d'intérêt en optimisation industrielle, et en particulier, l'optimisation robuste, multicritère, en présence d'incertitudes.
Le cours illustre ensuite la place de l'optimisation dans les principaux algorithmes d'apprentissage mathématique (machine learning). Ces questions sont illustrées par des exemples de problèmes de classification et de régression en apprentissage supervisé. Ces exemples sont l'occasion de discuter des questions de métriques et de procédures pour l'évaluation de l'apprentissage, de la validation et de l'inférence (crossfold, overfitting, etc).
Le cours présente les différentes classes d'apprentissage : non-supervisé, supervisé, par transfert, par renforcement, incrémental, etc. Les questions autour de la gestion des bases de données sont abordées : génération, imputation, visualisation, découpage. Le cours présente les liens entre l'apprentissage mathématique par transfert (transfer learning) et la simulation numérique pour adresser les questions de génération de bases de données synthétiques, d'imputation, de prédiction non-intrusive, d'inférence rapide, etc.
Le cours comporte une partie importante de projets informatiques au fil de l'eau. Toutes les séances ont lieu en environnement informatisé et permettent une mise en oeuvre immédiate des éléments théoriques.

Objectifs : Faire le lien entre optimisation numérique et apprentissage mathématique. Découvrir le machine learning au travers d'exemples concrets.

Pré-requis nécessaires : Bases d'analyse, solutions numériques des équations différentielles ordinaires, algèbre linéaire numérique, expériences en programmation en langage interprété.

Pré-requis recommandés : Cours L3 semestre 2 d'optimisation. Programmation en Python

UE HAX712X – Développement logiciel

Description : Ce cours est axé sur la découverte des bonnes pratiques de codage pour un niveau professionnel.
Le langage utilisé est Python, mais certains éléments de bash et de git seront également utiles. Un accent particulier sera mis sur le traitement et la visualisation des données au cœur du cours. Nous nous concentrerons principalement sur les concepts de base de la programmation, ainsi que sur la découverte des bibliothèques scientifiques de Python, dont "numpy, scipy, pandas, matplotlib, seaborn".

Au-delà des connaissances de ces packages fondamentaux, nous introduirons des pratiques modernes pour le code : tests (unitaires), contrôle de version (git), génération de documentation automatique, etc.

Objectifs : Être capable de créer un package de niveau professionnel en Python incluant, versionnement du code, test unitaire et documentation.

Pré-requis nécessaires : Les étudiants doivent connaître les notions de base des probabilités, de l'optimisation, de l'algèbre linéaire et des statistiques. Un minimum de connaissance des structures de base de la programmation est aussi demandé (if, then... else, boucle while/for).

Pré-**requis recommandés :** Un minimum de connaissance en Python serait un plus, ainsi qu'en analyse numérique (floatant, erreurs d'arrondis, etc.).

Programme :

1. [codage : algorithmes, modules, types de base, fonctions, boucles](Intro-Python/)
2. [codage : liste, dictionnaire, tuples, if et boucles, exceptions](Intro-Python/)
3. [classes (`__init__`, `__call__`, etc...), surcharge des opérateurs, gestion des fichiers](Intro-Python/), [git : une première introduction](Git/)
4. [`numpy` : notions de base sur les matrices (arrays), le découpage, l'algèbre linéaire simple, le masquage ; `matplotlib` : premiers tracés](Numpy-Matplotlib/)
5. [`github`, création d'une clé ssh, diverses commandes git, conflit, pull request](Git/)
6. [`numpy` : casting, concaténation, `imshow`, `meshgrid`, casting, copy](Numpy-Matplotlib/) ; [`scipy` : EDO, Interpolation, Optimisation](Scipy/)
7. [hands on git](Git/), [intro aux bases de linux et aux outils en ligne de commande](Bash)
8. [`scipy` : Images/canal, FFT](Scipy/), [Pandas : données manquantes](Pandas/)
9. [bash, regexp, grep, find, rename](Bash/), [environnement virtuel Python](Venv/)
10. [Pandas : premiers pas](Pandas/)
11. [Python virtual env : Anaconda](Venv/), [IDE : VScode](IDE/), [Create a Python Module](Python-modules/)
12. [Pandas : en savoir plus](Pandas/)
13. [Créer un module Python](Python-modules/), [tests unitaires](Tests-CI/)
14. [test unitaire](Tests-CI/)
15. [Matrices et graphiques éparses et mémoire](TempsMemoire/)
16. Numba
17. [Documentation avec Sphinx](Docs/)
18. Statsmodels

UE HAX713X – Théorie de l'information et de la décision

Description : La modélisation statistique est fondée sur les deux notions fondamentales d'information (qu'il s'agit d'extraire des données) et de décision (qu'il s'agit de prendre au vu de ces données). Ce cours introduit à la formalisation théorique de ces deux notions. Il est donc logiquement placé en début de cursus, beaucoup d'autres cours utilisant ses notions et résultats par la suite.

Introduction

Environnement aléatoire: la problématique de la réduction de l'incertitude et des risques associés.

I - Théorie de l'information

1. Entropie d'une distribution

a) Repérage d'un élément dans un ensemble probabilisé. Codage optimal et entropie d'une variable discrète. b) Entropie d'une variable continue. c) Entropie d'un vecteur. d) Propriétés générales de l'entropie: changement de variable affine, indépendance.

2. Information mutuelle

a) Information mutuelle et entropie conditionnelle de deux événements. b) Information mutuelle et entropie conditionnelle de deux variables. c) Contraste de Kullback-Leibler. Approximation du chi2. Application à l'information mutuelle de deux variables.

3. Esquisses d'applications statistiques

a) Sélection de variables prédictives. b) Arbres de régression et classement (CART). c) Classification par arbres de segmentation. d) Sélection de loi dans un modèle paramétrique. Pseudo-vraie loi. e) Recodages optimaux d'une variable: cas non-supervisé / supervisé.

II - Théorie de la décision

1. Cadre et problématique.

a) Expérience aléatoire, état de la nature, décision, perte, règle de décision pure / mixte, risque. b) Pré-ordre sur les règles de décision.

2. Quelques problèmes classiques:

a) Estimation ponctuelle ; b) Estimation ensembliste. ; c) Tests d'hypothèses. d) Diagnostic (classement).

3. Le pré-ordre sur les règles de décision.

a) Absence générale de règle optimale: absence d'estimateur optimal, absence de test optimal ; b) Règles admissibles ; c) Classe de règles essentiellement complète. Théorème de convexité.

4. Principes statistiques & sélection de règles

a) Principes permettant de transformer le préordre partiel en préordre total: principe minimax, principe de Bayes ; b) Principes de sélection: règles sans biais, règles fondées sur la théorie de l'information.

5. Approfondissement du cadre bayésien

a) Densité a priori (prior) du paramètre. Densité jointe de paramètre et des observations. Densité a posteriori du paramètre ; b) Quels Priors? Priors conjugués. Priors non-informatifs: prior uniforme, prior de Jeffrey ; c) Risque de Bayes. Règle de Bayes: estimateur de Bayes et test de Bayes. Intervalle de crédibilité.

Objectifs : Introduire les deux notions qui sont au fondement de la statistique mathématique: l'information (quantification de l'information et codage) et la décision (quantification et gestion du risque).

Pré-requis nécessaires : Cours de théorie des probabilités.

Pré-requis recommandés : une bonne maîtrise du calcul des probabilités, de la dérivation et de l'intégration.

UE HAX814X – Modèle Linéaire

Description : Le modèle linéaire est un outil à la fois simple et très riche, qui est à la base de nombreuses méthodes statistiques. Sa maîtrise et sa bonne compréhension sont très utiles à la fois d'un point de vue pratique, pour analyser de manière fine certains jeux de données, et d'un point de vue conceptuel, pour comprendre les bases théoriques des méthodes d'apprentissages plus avancées, y compris actuelles.

Ce cours propose une introduction au modèle linéaire de régression simple et multiple, avec des variables quantitatives ou qualitatives. Il présente la dérivation formelle et l'étude théorique des estimateurs des moindres carrés et du maximum de vraisemblance dans le cas gaussien. Il donne également des outils de validation et de sélection de variables pour étudier les limites du modèle. Enfin, il introduit l'utilisation pratique de cet outil sur des jeux de données simples grâce au logiciel R.

Objectifs :

- Comprendre les propriétés théoriques des modèles linéaires, avec variables quantitatives ou qualitatives.
- Savoir construire et estimer un modèle linéaire sur des données avec le logiciel R.
- Être capable d'interpréter les résultats et les limites du modèle posé.

Pré-requis nécessaires :

- Probabilités et statistiques descriptives de niveau L
- HA707X Statistique inférentielle
- Algèbre linéaire de niveau L

Pré-requis recommandés :

- La connaissance du logiciel R serait un atout

Programme :

1. Régression Linéaire Simple
   - Moindres Carrés : Propriétés des estimateurs des moindres carrés, prédictions, interprétation géométrique
   - Modèle Gaussien : Loi des estimateurs du maximum de vraisemblance, intervalles et régions de confiance
2. Régression Linéaire Multiple
   - Moindres Carrés : Propriétés des estimateurs des moindres carrés, prédictions, interprétation géométrique
   - Modèle Gaussien : Loi des estimateurs du maximum de vraisemblance, intervalles et régions de confiance
   - Sélection de Variable : Tests d'hypothèses, critères de sélection de modèle (AIC, BIC, ...)
3. Validation de Modèle
   - Analyse des résidus : Structure, normalité, homoscédasticité
   - Points leviers et données aberrantes : Matrice de projection, distance de Cook
   - Régression Linéaire avec variables qualitatives
4. ANOVA à un facteur
   - Modèle, tests d'hypothèses
   - ANOVA à deux facteurs
   - Modèle, tests d'hypothèses

UE HAX818X – Séries temporelles

Description : Ce cours d'introduction aux séries temporelles, c'est-à-dire une suite d'observations réalisées au cours du temps, constitue une boîte à outils indispensable pour le traitement de ce type de données fréquemment rencontrées dans un grand nombre

d'application : concentration d'un polluant dans l'air au cours du temps, taux de glucose dans le sang au cours du temps, ventes d'un produit dans une grande surface, cours d'une action à la bourse, etc. Ce cours s'attache à la fois à la présentation mathématique des concepts et aux aspects plus techniques de la mise en oeuvre des méthodes. Les illustrations numériques sont proposées avec le logiciel R.

Objectifs : maîtriser les principaux concepts pour la modélisation des séries temporelles. Etre capable de proposer une méthode adaptée de modélisation et de prédiction d'une série temporelle.

Pré-requis nécessaires : analyse, probabilité et statistique de niveau L3.

Pré-requis recommandés : Connaissance basique du Logiciel R

Programme :

1- analyse descriptive d'une série temporelle
2- processus ARMA, auto-corrélogrammes et auto-corrélogrammes partiels.
3- analyse spectrale
4- Prédiction linéaire : équations de Yule-Walker, algorithme de Durbin-Watson
5- estimation
6- test du portmanteau

UE HAX809X – Estimation et tests non-paramétriques

Description : Les méthodes non-paramétriques sont importantes dans de nombreuses applications statistiques car elles permettent de s'affranchir des approches classiques qui demandent la spécification de modèles statistiques valides. Or établir la validité d'un tel modèle est une entreprise complexe.

Les méthodes non-paramétriques contournent ce problème en utilisant la transformation des données en rang et en conditionnant sur certains quantités issues de la configuration observée de ces rangs. Les statistiques ainsi construites sont indépendantes de la loi des données brutes ce qui permet de construire des procédures d'inférence statistique libre du modèle sous-jacent aux données. En outre, la perte d'efficience statistique est minime.

Ce cours constitue une présentation assez complète des méthodes non-paramétriques. Il s'inscrit dans le prolongement d'un premier cours d'introduction aux méthodes inférentielles paramétriques, en adaptant et développant la théorie de plusieurs concepts avancés comme les tests conditionnels, la puissance comparée des tests (mesures d'efficacité), la notion de « effect size ». Il met l'accent sur l'application pratique de ces méthodologies en faisant un tour

d’horizon des principales commandes R et de leurs utilisations.

Objectifs : Amener l’étudiant à comprendre les limitations des méthodes statistiques paramétriques, à choisir face à un problème donné une bonne approche non-paramétrique en s’appuyant sur les principes statistiques sous-jacents, à mettre en œuvre via le logiciel R la solution de son problème et à rapporter aux utilisateurs finaux les conclusions de ses analyses et leur portée

Pré-requis nécessaires : cours de statistique inférentielle HAX707X de niveau M, cours de probabilité de niveau L

Pré-requis recommandés : cours de statistique descriptive de niveau L

Programme :

1. Définitions de *Statistique non paramétrique*

2. Deux astuces pour enlever la dépendance sur un paramètre inconnu : le conditionnement (avec application aux tables de contingence, tests du chi-deux et tests de Fisher-Yates) et la transformation en rang.

3. Le test de Wilcoxon-Mann-Whitney pour 2 échantillons : Les hypothèses et la statistique du test ; Comportement exact et asymptotique sous H0, le cas de données ex-aequo, robustesse du test.

4. Variantes et extensions du test de Wilcoxon-Mann-Whitney : Estimation ponctuelle et par intervalle de l’effet des traitements, le cas d’échantillons appariés, puissance du test de Wilcoxon-Mann-Whitney i) quand le paradigme de Student tient et ii) ne tient pas. Notions de « effect-size », calcul de tailles d’échantillons pour obtenir une puissance visée.

5. Autres tests pour le cas de 2 échantillons : test de Kolmogorov-Smirnov, test de Ansari-Bradley

6. Le cas de K >2 échantillons : test de Kruskal-Wallis, test de Friedman-Tukey. Le problème des comparaisons multiples. Contrôle de la p-value : méthode de Holm, méthode FDR (false discovery rate)

7. Indépendance, corrélation et régression : Coefficient de corrélation de Pearson, Spearman, Kruskal, test d’indépendance. Application à la régression

Mise en oeuvre avec le logiciel R des principaux tests non-paramétriques vus dans le cours.

UE HAX806X – Contrôle stochastique

Description : Ce cours constitue une introduction au contrôle stochastique. Dans ce type de problèmes, on cherche à modifier la trajectoire naturelle d'un processus pour remplir un certain objectif. Nous nous placerons dans le cadre des processus de Markov décisionnels à temps discret où on peut choisir une action à chaque pas de temps. Nous verrons comment formaliser les problèmes de contrôle stochastique dans ce cadre, et comment les résoudre théoriquement et numériquement.

Objectifs :

Savoir modéliser un problème de contrôle stochastique sous forme de processus markovien décisionnel

Savoir mettre en œuvre l'algorithme de programmation dynamique pour calculer les performances et stratégies optimales.

Pré-requis nécessaires :

Cours de processus stochastique de M1 (chaînes de Markov)

Vecteurs gaussiens

Logiciels scientifiques (R)

Pré-requis recommandés : Théorie de la mesure

UE HAX815X – Programmation R

Description : ce cours de programmation R s'adresse aux étudiants qui auront dans leur pratique professionnelle de connaître un langage de programmation pour faire du traitement avancé de données. Il s'agit donc d'apprendre à structurer, commenter et débugger proprement un code. Ce cours s'adresse à la fois aux étudiants du M1 SSD et du M1 Bio-Info. Il n'est pas destiné à utiliser les packages comme boîtes noires pour la mise en oeuvre de méthodes statistiques.

Objectifs : Maîtriser la syntaxe de base du langage r pour faire d la programmation scientifique.

Pré-requis recommandés : cours de statistique et de probabilité de niveau L. Connaître un langage de programmation serait un plus.

Programme :

1. Importation/Exportation de données
2. Graphiques
3. La programmation:

- algorithmes, modules, tests.
- variables, vecteurs, tableaux, listes.
- fonctions, conditions, boucles.
- objets
- interprétation & compilation
- librairies dynamiques

4. Principes et pratique du débogage.
5. Gestion des versions
6. Documentation
7. Environnements de développement intégré.

UE HAX817X – Projet

Description : Projets tutorés réalisés en groupe sous la direction d'un enseignant-e.
Objectifs : Mener une recherche bibliographique sur un sujet théorique ou appliqué. Travail en équipe. Rédaction d'un rapport écrit. Présentation orale.

UE Optionnelles (à choix)

UE HAU801I – Alignement et Phylogénie (4 ECTS)

Description : Ce module est composé de 2 parties. La première partie vise à comprendre les principes et méthodes d'alignement de séquences génétiques et de savoir utiliser les logiciels qui les implémentent : alignement de séquences 2 à 2 (dotplot, heuristiques BLAST et FASTA), utilisation des matrices de scores et des pénalités de gap, alignement multiple (global et local) et HMM (Hidden Markov Model). La deuxième partie se focalise sur la phylogénie et les méthodes de reconstruction d'arbres phylogénétique (basées sur les distances, la Parcimonie et les méthodes probabilistes).

Objectifs : Comprendre les méthodes d'alignements de séquences 2 à 2 et multiples, la construction des matrices de scores et le principe des HMM (Hidden Markov Model). Savoir utiliser les outils qui implémentent ces méthodes d'alignements. Comprendre les principales

notions de phylogénie ainsi que les méthodes de reconstruction d'arbres phylogénétiques. Savoir utiliser les outils correspondants.

UE HAX807X – Economie Générale (4 ECTS)

Description : Ce cours a pour objectifs l'acquisition de concepts de base en économie et l'approfondissement des connaissances ainsi acquises dans les domaines monétaires et financiers.
Il s'agira de comprendre la nature des interrelations entre les économies et d'analyser les conditions d'efficacité des politiques économiques en économie ouverte avec prise en compte de la nature du régime de change d'une part et du degré d'ouverture des capitaux d'autre part.
Dans cette perspective, la balance des paiements sera présentée et analysée ; les questions de compétitivité et d'attractivité des économies seront discutées.
Il s'agira également d'appréhender les impacts de la libéralisation financière sur la volatilité des changes et les possibilités qui s'offrent à différents types d'acteurs économiques pour se couvrir contre le risque de change.
Enfin, les notions de crise seront présentées (financière, de change). L'endogénéité des crises sera mise en avant avec une analyse des deux dernières crises : la crise financière de 2008 d'une part et la crise sanitaire d'autre part.

Programme :

CHAPITRE 1 LE SYSTÈME ECONOMIQUE (Circuit économique, agents économiques, identités comptables)
- Partie 1 : Union européenne et Zone euro
- Partie 2 : Introduction à la macroéconomie

CHAPITRE 2 LA POLITIQUE ECONOMIQUE
- Partie 1 : La logique de conception de la politique économique
- Partie 2 : Les raisons de l'intervention de l'Etat dans l 'économie
- Partie 3 : Les objectifs de la politique économique

CHAPITRE 3 LA BALANCE DES PAIEMENTS
- Partie 1 : Définitions et règles d'enregistrement
- Partie 2 : Les soldes significatifs de la balance des paiements
- Partie 3 : Les notions de compétitivité et d'attractivité

CHAPITRE 4 LES THÉORIES EXPLICATIVES DU TAUX DE CHANGE
- Partie 1 : Les approches réelles du taux de change
- Partie 2 : Les approches financières du taux de change
- Partie 3 : Instabilité du taux de change et comportement mimétique des agents

CHAPITRE 5 LA COUVERTURE CONTRE LE RISQUE DE CHANGE
Partie 1 : La notion de risque de change
Partie 2 : Les techniques internes de couverture contre le risque de change
Partie 3 : Les techniques externes de couverture contre le risque de change

UE HAU804I – Information biologique (2 ECTS)

Description : Systèmes NOSQL transactionnels : graphe/colonne/document, mécanismes de persistance et distribution de données volumineuses dans le contexte des sciences du vivant. Les ontologies terminologiques (OBO) et leurs usages dans la recherche d'information sont également abordés dans le module.

Objectifs : Comprendre les grands principes de la gestion de données volumineuses, complexes, distribuées et évolutives.

UE HAU802I – Bioinformatics Learning Lab – BILL (2 ECTS)

Analyses communes des données de séquençages issus de TP de biologie moléculaire.

UE HAR803B – Outils d'épidémiologie (2 ECTS)

Description : Différents « outils » et méthodes utilisés en épidémiologie sont abordés permettant une première initiation et l'acquisition d'éléments de base quelle que soit la spécialisation poursuivie (recherche ou gestion). Ces méthodes concernent aussi bien techniques liées à la manipulation et l'analyse biologique des agents pathogènes, que les méthodes d'enquêtes et les structures de la surveillance.

Objectifs : Connaître les bases de l'identification et de la gestion des risques biologiques liés à la manipulation d'agents pathogènes. Comprendre les principes techniques de réalisation de tests diagnostiques ainsi que leurs limites. Comprendre le fonctionnement des plateformes de traitement de matériel génétique et génomique. Savoir calculer et interpréter les indices d'épidémiologie descriptives. Connaître les bases des méthodes d'enquête en sciences sociales. Comprendre la structuration de la surveillance en santé.

UE HAX813X – Micro-économie (2 ECTS)

Description : Ce cours introduit à la modélisation mathématique du comportement d'acteurs cherchant à optimiser un objectif individuel en situation concurrentielle.

1. Introduction :
    a. Evolution de la définition de la science économique ; b. Rapide présentation de l'histoire de la Théorie des Jeux.

2. Interaction entre acteurs économiques et éléments de théorie des jeux non-coopératifs :
    a. Description d'un jeu

- Jeux simultanés et jeux dynamiques, Représentation d'un jeu en forme normale, Représentation en forme extensive
- Equilibre en stratégies dominantes.
- Equilibre de Nash
- Exemple d'application de modèle à l'économie industrielle : Duopole à la Cournot et à la Bertrand.

    b. Jeux dynamiques

- Raffinement de l'équilibre de Nash : équilibre parfaits en sous-jeux et algorithme de résolution vers l'amont (backward induction)
- Jeux d'entrée sur le marché

3. Théorie de l'information: asymétrie de l'information et incitations:
    a. Modèle principal-agent

- Asymétrie de l'information : Modèle des « Lemons » d'Akerlof
- Modèle d'entrée sur le marché en présence d'asymétrie d'informations sur les coûts.

    b. Incitations et mécanismes

- Design des mécanismes d'incitation
- Concrétisation en stratégie dominante.

Objectifs : Introduire le formalisme de la maximisation de l'utilité pour un acteur, puis deux en interaction. Établir les principaux résultats d'optimalité et appliquer ce formalisme au contexte économique concurrentiel.

Pré-requis nécessaires : Cours d'optimisation convexe.

Pré-requis recommandés : une bonne maîtrise du calcul différentiel et intégral, ainsi que de la résolution de problèmes d'optimisation sous contraintes.

UE de la deuxième année de MASTER SSD

UE HAX904X – Analyse Multivariée

Description :

La taille des données statistique ne cesse de croître, et notamment la richesse de la description des unités statistiques. Or, la modélisation statistique linéaire classique devient invalide en grande dimension, c'est-à-dire lorsque le nombre des variables dépasse celui des unités statistiques. Ce cours présente les techniques les plus courantes utilisées pour régulariser les modèles linéaires en grande dimension.

Introduction

Données de grande dimension. Réduction dimensionnelle et régularisation.

I - Modélisation linéaire régularisée d'une variable continue.

1. Le modèle linéaire classique.

a) Rappels express. b) Les pannes dues aux colinéarités.

2. Régression sur composantes principales.

a) La méthode. b) Qualités et défauts.

3. Régression PLS.

a) Critère et programme de rang 1. b) Critère et programme de rangs ultérieurs. c) Pourquoi PLS régularise. d) Choix du nombre de composantes pour la prédiction. e) Métrique du continuum entre OLS et PLS.

4. Régressions linéaires pénalisées.

a) Régression ridge. b) LASSO. c) Elastic net.

II - Modélisation linéaire régularisée d'un groupe de variables continues.

1. Le modèle linéaire gaussien multivarié

a) Le modèle classique. b) Le modèle pénalisé. c) Le modèle de MANOVA.

2. Régression PLS multivariée.

a) Critère et programme de rang 1 avec métriques quelconques. b) Cas particuliers: analyse canonique, ACP sur Variables Instrumentales, Régression PLS2. c) Critère et

programme de rangs ultérieurs. d) Prédiction: choix du nombre optimal de composantes. e) Métriques du continuum entre Analyse Canonique, ACPVI et PLS.

III - Modélisation linéaire d'une variable nominale: analyses discriminantes linéaires.

1. Analyse factorielle discriminante

a) Critère et programme. b) Composantes et pouvoirs discriminants.

2. Analyse discriminante PLS.

a) Critère et programme. b) Composantes et pouvoirs discriminants. c) Analyse discriminante barycentrique.

3. Aspects décisionnels.

a) Décision (classement), pertes, règles de décision (affectation), risques. b) Choix du bon nombre de composantes pour la décision.

Objectifs : Former à la modélisation linéaire uni- et multi-variée en grande dimension, c'est-à-dire à diverses techniques de régularisation de la modélisation linéaire classique.

Pré-requis nécessaires : Cours d'analyse des données multidimensionnelle (ACP & CA). Cours de géométrie euclidienne, d'espaces vectoriels normés et de réduction des endomorphismes.

Pré-requis recommandés : Cours de statistique descriptive univariée et bivariée. Bonne maîtrise du calcul matriciel.

UE HAX907X – Apprentissage Statistique

Description : Ce cours traite du cadre de l'apprentissage automatique sous un angle statistique.
Nous nous intéresserons principalement au cadre supervisé (régression et classification) et introduirons quelques éléments du cadre non-supervisé à travers les méthodes de partitionnement (clustering).
Au-delà des aspects de modélisation et de théorie, le cours couvrira aussi quelques éléments d'optimisation et d'implémentation (sklearn, pytorch, etc.) des méthodes introduites.

Objectifs : Être capable de modéliser un nouveau problème d'apprentissage au vu des objectifs, méthodes disponibles.

Pré-requis nécessaires :

- Modèle linéaire (HAX814X)

- Développement logiciel (HAX712X)
- Optimisation (HAX706X)

Pré-requis recommandés : Statistique inférentielle (HAX710X), Estimation et tests non paramétriques (HAX809X)

Programme :

1. Introduction à l'apprentissage supervisé; modèles linéaires.
2. Validation croisée, régression logistique, analyse discriminante.
3. Sélection de modèle et méthodes de régularisation.
4. Mesure de performance (multi-classe: top-k, ROC curve AUC, etc.)
5. Perceptron et (descente) de gradient stochastique.
6. SVM
7. Arbres de décision, Forêts aléatoires et Boosting.
8. Apprentissage non-supervisé (partitionnement: KMEANS, Ward's method)
9. Résaux de neurones

UE HAX912X – Modèles Linéaires Généralisés

Description : ce cours introduit le cadre général des modèles linéaires où l'on cherche à exprimer une variable réponse en fonction d'une combinaison linéaire de prédicteurs. En faisant l'hypothèse à la fois une relation spécifique entre la réponse moyenne et les prédicteurs (fonction de lien), ainsi qu'une distribution spécifique de la variation aléatoire de la réponse autour de sa moyenne, il est possible de représenter des données binaires (ex.: présence/absence, mortalité/survie) ou de comptage (ex.: nombre d'individus, nombre d'espèces). Grâce à ce cadre général, on peut alors modéliser des variables non-normalement distribuées. L'utilisation et l'interprétation des modèles de régression logistique, binomiale et de Poisson seront en particulier détaillés.

Objectifs : Etre capable de modéliser la relation entre une variable réponse qu'elle soit continue, discrète ou catégorielle. Savoir mettre en oeuvre une méthode numérique pour l'estimation, tests, diagnostics, savoir comparer et choisir un modèle dans le cadre d'un GLM.

Pré-requis nécessaires : probabilités de niveau L, modèle linéaire, statistique inférentielle

Programme :

Introduction : rappels sur le modèle linéaire (gaussien)

1. Famille exponentielle : définition et propriétés
2. Prédicteurs linéaires et fonctions de lien classiques : identité, logit, logarithme
3. Estimation : équations de vraisemblance, scores de Fisher
4. Modèle logistique et discrimination
5. Modèles de comptage : binomial et Poisson
6. Modèles à sur- et sous-dispersion

UE HAX916X – Projets ou soutenance alternance

UE HAX906L – Anglais

Description : Cours TD d'anglais, à l'intention des étudiants de la filière « M2 Stat et sciences des données » et qui visent l'autonomie professionnelle en langue anglaise.

Objectifs : Renforcer et consolider les bases linguistiques selon les 5 compétences langagières décrites par le Cadre Européen Commun de Références en Langues (CECRL). Permettre aux étudiants d'accéder à une aisance orale et écrite compatible avec le travail avec des interlocuteurs anglophones.

Contrôle continu intégral – La présence et une participation active aux cours seront exigées.

Programme :

- Compréhension orale – supports vidéo, échanges en groupe
- Compréhension écrite – à partir d'articles de la presse scientifique
- Expression orale en interaction – entretiens et travaux en groupe
- Expression écrite – comptes rendus de compréhension orale
- Expression orale en présentations individuelles

Pré-requis nécessaires :

Compréhension écrite et orale, notions de grammaire et compétences d'expression écrite et orale élémentaires.

Pré-requis recommandés :

Le niveau B2 du CECRL à l'oral comme à l'écrit

UE HAX908X – Estimation non-paramétrique

(Cours proposé dans le parcours SSD – BIOSTAT uniquement)

Description : Ce cours présente quelques-unes des méthodes classiques et modernes pour la construction d'estimateurs non-paramétriques de la densité ou de la régression. Des aspects tout à la fois théoriques et pratiques y sont abordés.

Objectifs : L'objectif de ce cours est double : d'une part il s'agit de comprendre la construction des estimateurs afin d'être capable de les adapter à d'autres contextes d'estimation et de comprendre les résultats mathématiques qui garantissent le bien-fondé de ces approches mais aussi les limites notamment en grande dimension. Un deuxième objectif est d'appréhender les enjeux de la sélection des paramètres d'un point de vue pratique grâce à la mise en oeuvre numérique des algorithmes pour la sélection de fenêtre ou la sélection de modèles. A l'issue de ce cours, l'étudiant doit disposer d'une boîte à outils pour l'implémentation pratique de ces méthodes.

Pré-requis nécessaires : cours d'analyse et de probabilités de licence,

Pré-requis recommandés : connaître un langage de programmation, HAX710X cours de statistique inférentielle, HAX814X cours de Modèle Linéaire

Programme :

1. Notions de risque et de critères d'erreur
2. Introduction à l'estimation non-paramétrique : la fonction de répartition empirique
3. Estimateurs à noyaux de la densité :
    - lemme de Bochner, risque quadratique
    - noyaux usuels, noyaux d'ordre supérieur
    - sélection de la fenêtre optimale : méthode plug-in, validation croisée, autres méthodes adaptatives
    - vitesse de convergence, comparaison avec la vitesse paramétrique
4. Estimateurs par projection de la densité :
    - bases de Fourier, bases de polynômes, ondelettes de Haar.
    - estimateur par minimum de contraste, étude du risque quadratique
    - sélection de la dimension du sous-espace de projection : méthode de Barron, Birgé, Massart (1999)
    - notion d'estimateur adaptatif (pour des modèles emboîtés).
5. Estimateurs de Nadaraya-Watson de la fonction de régression : approche par quotient.
6. Estimateurs des moindres carrés : lien avec le modèle linéaire multivarié.

- sélection de modèles, adaptation.

7. Régression polynomiale locale : mise en oeuvre pratique des splines
8. Fléau de la dimension pour estimer une densité multivariée et/ou une fonction de régression de plusieurs variables. Visualisation en dimension 2 : exemple de données géo-spatialisées. Quelques pistes pour la dimension supérieure à 1 : modèles single-index, modèles additifs.
9. Bootstrap : construction d'intervalles de confiance

UE HAX918X – Statistique Bayésienne

(Cours proposé dans le parcours SSD – BIOSTAT uniquement)

Description : Ce cours propose une introduction à la statistique bayésienne paramétrique. Après la présentation du paradigme bayésien les cas des estimations ponctuelles et ensemblistes seront considérés puis la méthodologie de choix bayésien de modèles sera abordée. Les modèles binomiaux, gaussiens et linéaires serviront d'illustration pour les thèmes précédents.

Pour les modèles complexes, les problématiques d'estimation et de sélection de modèle dans le contexte bayésien nécessitent le recours à des outils évolués d'approximation d'intégrales. Aussi, la deuxième partie du cours sera centrée sur les méthodes de Monte Carlo et les algorithmes de Monte Carlo par Chaînes de Markov.

Objectifs : Fournir les principaux outils de la statistique bayésienne. Etre capable de les mettre en oeuvre numériquement.

Pré-requis nécessaires : cours de probabilités et statistique inférentielle de niveau M1.

UE AAMSD201 – Management des risques

(Cours proposé dans le parcours SSD – MIND uniquement)

UE HAX006X – Modèles à variables latentes

Description : Beaucoup de phénomènes ne sont qu'incomplètement ou indirectement observés, ce qui complique leur analyse. Leur modélisation statistique doit alors inclure des variables non observées, dites latentes, qui sont rattachées d'une façon ou d'une autre aux variables observées. Ce cours introduit aux diverses manières d'introduire des variables latentes dans un modèle selon leur type (qualitatives ou quantitatives), et de procéder à l'estimation des paramètres du modèle.

Objectifs : Former à la modélisation statistique en présence de variables non observées ou indirectement observées.

Pré-requis nécessaires : Cours d'analyse des données multidimensionnelle (ACP & CA). Cours d'analyse multivariée. Cours de statistique inférentielle.

Pré-requis recommandés : Très bonne maîtrise de l'algèbre matricielle, de la dérivation vectorielle et de l'optimisation sous contraintes.

Programme :

Introduction. Situations et typologie de variables latentes: continues / nominales; aléatoires / non-aléatoires.

I - Modèles à VL non aléatoires

1. VL nominale: modèles de classification (clustering)

    a) Modèle gaussien. b) Modèle multinomial.

2. VL continues: modèles à composantes

    a) Modèle d'ACP. b) Modèle d'ACPVI. c) Modèle PLS. d) Modèles explicatifs multi-blocs: THEME et SCGLR.

II - Modèles à VL aléatoires

1. L'algorithme EM
2. VL nominale: modèles de mélange

    a) Mélange gaussien. b) Mélange multinomial: latent class analysis.

3. VL Continues: modèles à facteurs

    a) Modèle à facteur pour 1 bloc. b) Modèles à équations structurelles.

4. Modèles à chaîne de Markov cachée

UE HAX002X – Analyse des durées de Vie

Description : La durée de vie d'un individu en biostatistique, ou d'une composante en analyse de la fiabilité, est une quantité dont l'analyse statistique se distingue des données habituelles. D'une part, elle mène à considérer des quantités comme la fonction de hasard, le temps moyen résiduel de vie, etc. qui n'ont pas autant d'intérêt dans d'autres domaines de la statistique. D'autre part, elle fait souvent intervenir un mécanisme de censure, du fait que les données sont observées de façon incomplète en raison de la longueur des expériences par rapport au temps qu'on veut leur allouer. Ce cours présente les bases de l'analyse de survie. Les raisons d'être et les principaux mécanismes de censure de données sont abordés. Deux grands types d'approches statistiques sont présentés : l'approche paramétrique, qui malgré ses limitations, a souvent la faveur des utilisateurs, car "les paramètres parlent" et l'approche non-paramétrique qui permet de conforter et de compléter les analyses paramétriques en leur donnant une souplesse et une profondeur accrue quand les données sont nombreuses. Le module présente aussi différents modèles (modèle de Cox, du taux de panne accéléré, etc.) permettant de relier la survie à des facteurs explicatifs, ce qui permet de déterminer ceux pouvant impacter cette survie. Cette information est particulièrement utile dans un contexte sanitaire, car elle permet de personnaliser les projections de survie d'un individu.

La mise en oeuvre de ces méthodes se fera sur le logiciel R.

Objectifs : Au terme de ce module, l'étudiant devra savoir conduire de façon raisonnée l'analyse statistique de données de survie en choisissant l'approche adaptée aux particularités des données et au type de censure rencontré. Il devra savoir estimer les quantités inconnues et produire l'inférence pertinente. Il devra pouvoir faire le lien entre une série de facteurs et la survie étudiée et identifier ceux qui ont un impact sur celle-ci ainsi que la magnitude de cet effet. Enfin, il devra pouvoir mener de façon autonome les aspects computationnels nécessaires à la production d'énoncés inférentiels.

Pré-requis nécessaires : HAX710X, HAX814X, HAX815X, HAX912X

Pré-requis recommandés : HAX809X, HAX908X

Programme :

1. Introduction aux données de survie : origine, utilités et particularités; les différents types de censure
2. Approche paramétrique :
    - Les différentes lois paramétriques utilisées.
    - Estimation des paramètres en présence de censure, comportement des estimateurs et production d'inférences : intervalle de confiance et test d'hypothèses pour une expérience de survie,
    - comparaison de plusieurs courbes de survie en utilisant le principe du maximum de vraisemblance. Modèles pour co-facteurs : introduction aux modèles de taux de panne accéléré, modèle de Cox.
    - Choix et validation de modèles, choix des variables explicatives importantes. Modèles de régression de survie accélérée. Utilisation des packages R adaptés.
3. Approche non-paramétrique :
    - Estimateur de Kaplan-Meier de la fonction de survie et estimateur de Breslow, estimateur de Nelson-Aalen de la fonction de risque cumulé,
    - construction des intervalles de confiance et formule de Greenwood, autres intervalles de confiance par transformation monotone.
    - Estimateur actuariel.
    - Tests non-paramétriques de rangs pour la comparaison de plusieurs groupes. Mise en oeuvre des packages R usuels.

UE HAX004X – Complément 1

(Cours proposé dans le parcours SSD – BIOSTAT uniquement)

Description : Les cours de complément présentent des ouvertures vers des domaines plus spécialisés de la statistique et de la modélisation stochastique. Leur contenu est susceptible de changer d'une année sur l'autre. Les thèmes abordés pourront être les suivants :
- analyse des séquences biologiques : Modèles probabilistes d'évolution des séquences biologiques, Inférence des phylogénies, Modèles de Markov cachés pour la détection de motifs, Modèles graphiques et inférence de réseaux de régulation génique
- dynamique des populations : processus de naissance et mort (définitions, propriétés, comportement asymptotique, estimation des paramètres, simulation), approximations déterministes, stochastiques ou hybrides

- statistique biomédicale : Introduction aux données de la recherche clinique, aspects réglementaires et méthodologiques, Fonction de vraisemblance et applications aux données bio-médicales, Rappels sur les données de survie, modèles à risques compétitifs, test basé sur une U-statistique, Modèles d'analyse de données de fertilité, Diagnostic médical et courbes ROC comme application d'une U-statistique, Méta-analyses.

- statistique des extrêmes et applications à l'environnement :Théorie des valeurs extrêmes univariée et multivariée : loi des maximas et des dépassements de seuils élevés pour des variables et des vecteurs aléatoires, dépendances extrémales, estimation de quantiles extrêmes, étude du risque. Applications pour des données environnementales : pluie, hauteur de vagues, températures...

- statistique spatiale : Introduction des éléments fondamentaux de la prédiction spatiale et applications. Afin de couvrir un large panel des statistiques spatiales, ce cours pourra s'articuler autour de deux parties : les processus ponctuels et la géostatistique .

- modèles linéaires mixtes : Extension des modèles linéaires aux modèles linéaires mixtes. Estimation des paramètres d'effet fixe comme ceux de variance au sein de ces modèles. Mise en œuvre sur différents cas pratiques. Effets aléatoires dans les modèles linéaires généralisés.

Objectifs : UE d'ouverture vers des domaines plus spécialisés de la statistique et de la modélisation stochastique.

UE HAX005X – Complément 2

(Cours proposé dans le parcours SSD – BIOSTAT uniquement)

Cf. HAX004X – Complément 1

UE AAMSD209 – Stratégie et Gestion de projet

(Cours proposé dans le parcours SSD – MIND uniquement)

UE AAMSD210 – Data Mining et données manquantes

(Cours proposé dans le parcours SSD – MIND uniquement)