



Programme  
des enseignements de 3<sup>e</sup> année  
Filière Data Science en Santé & Biostatistique  
ANNÉE SCOLAIRE 2022 / 2023



École nationale  
de la statistique  
et de l'analyse  
de l'information

# FILIÈRE DATA SCIENCE EN SANTÉ & BIOSTATISTIQUE

ANNÉE SCOLAIRE 2022/2023

*HEALTH DATA SCIENCE & BIOSTATISTICS SPECIALIZATION*

*2022/2023 ACADEMIC YEAR*

# Table des matières

Présentation de la filière .....	4
Descriptifs des enseignements communs.....	11
UE : APPRENTISSAGE AUTOMATIQUE (MACHINE LEARNING) .....	12
APPRENTISSAGE AUTOMATIQUE .....	13
APPRENTISSAGE PROFOND .....	14
Descriptifs des enseignements de la filière .....	16
UE SPECIFIQUES FILIERE BIostat.....	17
UE 2 : Modèles de regression en biostatistique.....	18
PLANS D'EXPERIENCES .....	20
MODELES MIXTES .....	21
ANALYSE DE DONNEES DE SURVIE .....	22
UE 3 : Modélisation stochastique et statistique .....	24
STATISTIQUE DES PROCESSUS .....	26
TRAITEMENT DE LA DONNEE MANQUANTE .....	28
STATISTIQUE BAYESIENNE .....	30
META-ANALYSE .....	32
UE 4 : Essais cliniques .....	34
ESSAIS CLINIQUES.....	36
PHARMACOMETRIE .....	38
ECHELLE DE MESURE EN SANTE .....	40
UE 5 : Épidémiologie et Données -omiques.....	42
EPIDEMIOLOGIE .....	43
MODELISATION COMPARTIMENTALE .....	44
ANALYSE DES DONNEES OMIQUES .....	46
UE : PROJETS.....	48
PROJET METHODOLOGIQUE .....	49
PROJET DE FIN D'ETUDES .....	50
DATA CHALLENGE.....	51
ANGLAIS.....	52
UE : PROJET PROFESSIONNEL ET STAGES .....	53
DROIT DU TRAVAIL .....	54
Offre de Formation Par la Recherche .....	55

## Présentation de la filière

---

La troisième année du cycle ingénieur de l'ENSAI comporte six voies de spécialisation. La spécialité Data Science en Santé et Biostatistique est la spécialité préparant aux métiers de la biostatistique. La biostatistique concerne les méthodes statistiques pour l'analyse et l'interprétation des données biomédicales. Les biostatisticiens travaillent avec des équipes multidisciplinaires pour concevoir, analyser et résumer les données des sciences expérimentales. Les domaines d'application sont le domaine de la santé, la médecine, l'agriculture, la foresterie, les sciences de l'environnement et au-delà.

L'objectif de cette spécialité est de former des biostatisticiens, mais également des scientifiques des données. Le programme intègre une unité d'enseignement (UE) sur l'apprentissage automatique, deux UEs ayant pour objet d'approfondir des outils de modélisation statistique (les modèles mixtes, l'analyse de survie, les plans d'expérience, l'inférence bayésienne et les processus stochastiques). En outre les domaines importants de la biostatistique sont couverts comme l'épidémiologie, les essais cliniques et les données -omiques. Transversalement à ces unités d'enseignement, les applications en informatique (R, SAS, Python, JAGS, etc) sont omniprésentes.

Par ailleurs la formation vise à ce que l'étudiant soit opérationnel en anglais. Outre les cours en anglais préparant au TOEIC, les étudiants seront initiés à la rédaction de rapports d'étude et de présentation orale en anglais à l'occasion du projet méthodologique ou du projet Essais Cliniques, sans oublier certains enseignements pouvant être dispensés en anglais.

Des séminaires professionnels présentent la richesse des métiers offerts en biostatistique. Ils sont l'occasion d'une présentation, par des praticiens, des outils ou modèles statistiques utilisés dans les entreprises et instituts de recherche.

### Pré-requis

Lors du second semestre de deuxième année, les élèves ingénieurs sont invités à choisir des options qui sont des pré-requis pour plusieurs cours de la filière ou qui permettront d'acquérir des compétences recherchées dans le domaine de la biostatistique. Aussi les étudiants qui souhaitent intégrer la filière sont invités à choisir des options principalement en probabilité et statistique afin de consolider leurs bases, ainsi que des cours de programmation :

- Chaînes de Markov, Martingales ;
- Calcul Bayésien, Méthodes de Rééchantillonnage, Régression avancée ;
- C++, R avancé, RShiny, SAS.

### Débouchés

A l'issue de la formation, l'étudiant peut intégrer des postes de Biostatisticien, ou ingénieur d'études ou *Data Scientist* au sein de l'industrie pharmaceutique, agro-alimentaire, les hôpitaux, et les instituts de recherche privés et publiques.

Il est également possible pour les étudiants de poursuivre en doctorat. Il est à noter que la filière dispose de plusieurs conventions de bi-diplomation afin de compléter la formation des étudiants.



### Offre de Formation Par la Recherche (OFPR)

L'Ensaï offre la possibilité, aux élèves de 3<sup>ème</sup> année qui le souhaitent, de se préparer en vue d'une carrière de chercheur en entreprise au sein des services Recherche et Développement ou dans le secteur académique. Dans le cadre de l'option formation par la recherche (Ofpr), ces élèves bénéficient d'un aménagement de leur scolarité leur permettant de suivre au choix un des trois masters recherche suivants :

- Master Recherche, mention Mathématiques et applications, spécialité Statistique Mathématique ;
- Master Recherche, mention Santé Publique, spécialité Modélisation en Pharmacologie Clinique et Epidémiologie (MPCE) ;
- Master Recherche, mention Bio-informatique, spécialité Bio-informatique en santé.



## Programme

	Volume Présentiel	Volume Total	Crédits
<b>UE Apprentissage Automatique</b>	<b>66 h</b>	<b>130 h</b>	<b>5</b>
Apprentissage Automatique	30 h	64 h	2.5
Apprentissage Profond	12 h	24 h	1
Traitement automatique du langage	21 h	42 h	1.5
<b>UE Modèle de Régression en Biostatistique</b>	<b>69 h</b>	<b>146 h</b>	<b>5</b>
Plans d'expériences	18 h	36 h	1
Modèles mixtes	24 h	52 h	2
Modèles de durée	24 h	52 h	2
<b>UE Modélisation Statistique et Stochastique</b>	<b>72 h</b>	<b>139 h</b>	<b>5</b>
Statistique des processus	21 h	46 h	2
Traitement des données manquantes	15 h	30 h	1
Statistique bayésienne	18 h	33 h	1
Méta-analyse	18 h	30 h	1
<b>UE Essais cliniques</b>	<b>75 h</b>	<b>160 h</b>	<b>5</b>
Essais cliniques	42 h	84 h	3
Pharmacométrie	18 h	48 h	1
Échelle de Mesure en Santé	15 h	30 h	1
<b>UE Épidémiologie / Données -Omiques</b>	<b>63 h</b>	<b>132 h</b>	<b>5</b>
Épidémiologie	15 h	30 h	1,5
Modélisation compartimentale	12 h	30 h	1
Analyse de données Omiques	36 h	72 h	2.5
<b>UE Projets</b>	<b>58 h</b>	<b>150 h</b>	<b>5</b>
Projet méthodologique	10 h	75 h	3
Projet de fin d'étude (Données Omiques)	18 h	30 h	1
Anglais	30 h	45 h	1
<b>UE Projet professionnel et stages</b>	<b>-</b>	<b>-</b>	<b>30</b>
Stage 2A	2 à 3 mois		5
Stage 3A	20 semaines à 6 mois		25
Séminaires Professionnels (dont Droits des Entreprises)	30 h		-
<b>TOTAL</b>	<b>412 h</b>	<b>863 h</b>	<b>60</b>

# Présentation des UEs

## UE Apprentissage Automatique

### *Machine Learning*

ECTS : 5 -- Volume horaire : 130 heures

L'Apprentissage Automatique traite de l'analyse et du développement de modèles et d'algorithmes capables d'apprendre à partir des données afin d'effectuer une analyse prédictive. Il trouve de nombreuses applications: prédiction des cours basés à partir d'articles de presse en finance, détection de maladie par imagerie médicale en santé, recommandation de produits en marketing, compression d'images ou encore modèles de traitement du langage, toutes ces applications reposent sur les mêmes bases. Cette UE vise à présenter la mise en œuvre des techniques classiques, en développant un esprit critique sur leurs limites (sur-apprentissage, grande dimension, représentativité de l'échantillon) et en utilisant des données non structurées (texte, image...).

## UE Modèles de Régression en Biostatistique

### *Regression Models in Biostatistics*

ECTS : 5 -- Volume horaire : 140 heures

L'UE *Modèle de régression en Biostatistique* a pour objet des modèles de régression largement utilisés dans la littérature de recherche médicale. Ces modèles ont en commun de mettre en relation une variable d'intérêt avec des prédicteurs multiples : régression linéaire multiple pour une variable d'intérêt continue, régression logistique pour variable d'intérêt binaire et le modèle de risques proportionnels de Cox pour les résultats liés au temps jusqu'à l'événement. Ces modèles étudiés durant les deux premières années du cycle ingénieur à l'Ensaï sont ici approfondis et illustrés par des cas pratiques en biostatistique.

## UE Modélisation Stochastique et Statistique

### *Stochastic and Statistical Modelisation*

ECTS : 5 -- Volume horaire : 139 heures

L'UE *Modélisation stochastique et statistique* a pour objet l'étude de modèles fréquemment utilisés en Biostatistique. Elle couvre l'étude des processus stochastiques couramment utilisés en épidémiologie ou en génomique, la modélisation bayésienne applicable à des problèmes complexes d'analyse des données, notamment en méta-analyse, ainsi que la considération du traitement des données manquantes spécifique aux essais cliniques ou études épidémiologiques.

## UE Essais Cliniques

### *Clinical Trials*

ECTS : 5 -- Volume horaire : 160 heures

Le développement d'une molécule est un processus complexe qui dure entre 5 et 10 ans. Il comporte de nombreux jalons afin de montrer aux autorités que la molécule est efficace tout en étant suffisamment peu toxique pour l'homme. Les essais cliniques sont des étapes essentielles de ce développement et ont pour but l'évaluation de nouveaux traitements ou de nouvelles modalités de soins. L'UE *Essai Clinique* a pour but d'initier les étudiants à une lecture plus critique de la littérature sur les essais cliniques et fournira les outils pour concevoir leur propre protocole. Au travers de cette UE, ils seront également initiés à la Pharmacométrie et à la Psychométrie.

## UE Epidémiologie et Données -Omiques

### *Epidemiology & Omics Data*

ECTS : 5 -- Volume horaire : 132 heures

L'UE a pour but d'initier les étudiants aux approches multi-omiques et à la modélisation en épidémiologie. Les approches multi-omiques visent l'intégration de données quantitatives obtenues pour différentes molécules biologiques afin de comprendre leurs interactions et d'avoir une image plus complète du phénotype biologique. L'épidémiologie a pour objet l'étude des causes des maladies infectieuses et non infectieuses, de faire progresser et d'évaluer les méthodes de prévention des maladies et d'aider à planifier et à évaluer l'efficacité des programmes de santé publique.

## UE Projets *Projects*

ECTS : 5 -- Volume horaire : 150 heures

Les projets sont l'occasion pour les étudiants de se mettre dans une démarche de recherche et développement : diagnostique d'un problème nouveau, lecture de la littérature scientifique sur le sujet, résolution d'un problème en respectant un compromis entre les règles de l'art d'une part, et les contraintes humaines, financières et techniques de l'autre. Ils permettent par ailleurs aux élèves de mesurer l'utilité de toutes les notions acquises au cours des trois années de formation.

Selon les filières, la réalisation d'un Data Challenge complète ces cas d'études concrets, à travers la réalisation d'un projet sur un temps court et des contraintes spécifiques. Le cours d'anglais visent soit à préparer les étudiants à passer le TOEIC (niveau B2) ou soit à consolider/entretenir leurs niveaux.

## UE Projet Professionnels et Stages *Professional Project and Internships*

ECTS : 30

Cette UE a pour but de contribuer à la bonne insertion professionnelle des étudiants, notamment dans un contexte international. Les séminaires professionnels aideront les étudiants à s'orienter dans les choix futurs de leur carrière, notamment le stage. Le cours de droit du travail permet aux étudiants de prendre connaissance du contexte juridique de l'entreprise.

En rejoignant les équipes statistiques d'une entreprise, d'une administration ou d'un centre de recherche, l'élève ingénieur est invité à mettre en application les connaissances acquises au cours de sa scolarité. Le stage de fin d'études doit permettre à l'étudiant d'atteindre au moins deux objectifs :

- Un objectif technique : une mission est confiée à l'élève stagiaire et il doit, à l'aide de ses connaissances théoriques et ses aptitudes, tenter de répondre à cette commande en utilisant au mieux son environnement proche.
- Un objectif professionnel : l'élève stagiaire évolue dans un environnement professionnel et il doit mettre à profit cette période de stage pour parfaire sa connaissance du monde du travail, accroître ses capacités relationnelles et adopter une démarche d'insertion.

# Compétences / Acquis d'Apprentissage

Une fois diplômé, un étudiant titulaire d'un diplôme ingénieur avec la spécialisation *Sciences des Données en Santé et Biostatistique* doit avoir acquis les compétences suivantes

## A – Adaptation aux exigences propres de l'entreprise et de la société

- A1. Intégrer les enjeux de l'entreprise (dimension économique, respect de la qualité, compétitivité et productivité, exigences commerciales, intelligence économique) dans l'exercice des missions confiées
- A2. S'insérer dans la vie professionnelle : exercice de la responsabilité, esprit d'équipe, management de projets
- A3. Travailler dans un contexte international et multiculturel avec la maîtrise d'une ou plusieurs langues étrangères et ouverture culturelle associées.

## B - Management et sécurisation des données (niveau intermédiaire)

- B1. Décrire la réglementation autour des données (anonymisation, RGPD, sécurisation ...)
- B2. Mettre en place un protocole réglementaire de recueil des données
- B3. S'assurer de la sécurité du stockage et du transfert des données.
- B4. Interroger les bases de données relationnelles en utilisant SQL
- B5. Expliquer les avantages et les inconvénients des bases de données NoSQL.

## C – Analyse quantitatives des données au travers des statistiques descriptives et inférentielles dans le domaine de la santé (niveau avancé)

- C1. Mener une étude bibliographique sur les données à analyser
- C2. Identifier les principaux problèmes rencontrés en Biostatistique et les différentes approches permettant d'y résoudre
- C3. Programmer en R, Python et SAS
- C4. Manipuler des données avec un logiciel statistique ou langage informatique
- C5. Calculer des statistiques descriptives uni- et multivariées
- C6. Maîtriser la théorie sous-jacente à la statistique inférentielle (probabilité, estimation ponctuelle, intervalle de confiance, test statistique)
- C7. Sélectionner et interpréter les statistiques apportant de l'information compte tenu de la problématique.

## D- Conception des études en médecine ou en santé publique (niveau avancé)

- D1. Aider à la clarification de la question de recherche, des hypothèses et des limites de l'étude
- D2. Aider à la sélection de la population cible, des paramètres appropriés
- D3. Choisir le modèle expérimental adéquat pour un problème et une situation donnée
- D4. Calculer la taille de l'échantillon en tenant compte de l'objectif de l'étude, des critères d'évaluation (contrôles de l'erreur de type 1 et 2, de la variabilité des estimations qui soient cliniquement pertinent) et du plan expérimental de l'essai clinique.
- D5. Préparer une analyse statistique en accordant une attention particulière à l'analyse des sous-groupes, la multiplicité des mesures et les données manquantes.
- D6. Contribuer à la conception et à l'évaluation d'un protocole d'essais cliniques
- D7. Décrire les grandes théories psychométriques afin de prendre en compte l'évaluation de la qualité de vie dans un protocole d'essais cliniques.
- D8. Choisir un schéma d'étude adapté à l'investigation d'un phénomène épidémiologique.
- D9. Planifier une méta-analyse dans une revue systématique dans le domaine des essais cliniques.

## E- Mise en œuvre de méthodologies innovantes en Science des Données pour l'analyse des données médicales et de santé publique complexes (niveau avancé)

- E1. Démontrer la maîtrise des principes qui sous-tendent les méthodes statistiques couramment utilisées en Biostatistique afin de permettre l'appréciation de nouvelles méthodologies ou de développer des solutions spécifiques à de nouveaux problèmes.

E2. Conduire l'inférence à l'aide des méthodes statistiques appropriées pour analyser les données généralement rencontrées dans les applications médicales, y compris les données binaires, catégorielles, données psychométriques, données de survie, données répétées, données longitudinales, données hiérarchiques, données affectées par l'erreur de mesure, les revues systématiques (méta-analyse) et les données -omiques.

- E3. Développer un modèle pour l'évolution temporelle de la concentration et de la réponse du médicament dans diverses situations
- E4. Utiliser des outils statistiques et des algorithmes d'apprentissage automatique pour identifier les corrélations, les effets, les modèles et les tendances dans les données, et de formaliser les prédictions
- E5. Créer et mettre en œuvre des méthodologies innovantes et des algorithmes pour analyser des données massives provenant de différentes sources, notamment des données issues des technologies -omiques
- E6. Reconnaître les forces et les faiblesses des méthodes analytiques et des sources de données.

F- Communication des analyses statistiques à des publics dans les disciplines connexes à la biostatistique

- F1. Rédiger en français et en anglais des rapports écrits des résultats en prenant en compte les aspects métiers, ainsi que la comparaison de différentes méthodologies statistiques.
- F2. Contribuer à la rédaction d'un protocole d'essais cliniques en anglais.
- F3. Présenter oralement une démarche scientifique en français et en anglais.

G- Développement d'outils de visualisation (niveau intermédiaire)

- G1. Définir le cahier des charges.
- G2. Concevoir une interface graphique pour l'analyse et la visualisation des données.
- G3. Déployer un outil-logiciel.

# Descriptifs des enseignements communs

---

UE 1 – Machine Learning

# UE : APPRENTISSAGE AUTOMATIQUE (MACHINE LEARNING)

Correspondant de l'UE : Sébastien DA VEIGA  
 Nombre d'ECTS : 5

Volume horaire de travail élève : env. 130h  
 (enseignements + travail personnel)  
 Nombre d'heures d'enseignement : 66h

## Finalité de l'UE :

L'apprentissage automatique (machine-learning) est un paradigme essentiellement différent des approches statistiques exploratoires (statistiques au sens classique) ou explicatives (économétrie). Il vise un objectif de prédiction dans la continuité des méthodes d'apprentissage statistique supervisé introduites lors des premières années de la formation d'ingénieur. Largement utilisé dans l'ensemble des professions statistiques à l'heure actuelle (les métiers de la *data science*), l'apprentissage statistique est incontournable dans la formation de l'ingénieur statisticien et trouve de nombreuses applications: prédiction des cours basés à partir d'articles de presse en finance, détection de maladie par imagerie médicale en santé, recommandation de produits en marketing, compression d'images ou encore modèles de traitement du langage, toutes ces applications reposent sur les mêmes bases.

## Structuration de l'UE :

L'UE se compose de 3 matières : apprentissage statistique (*machine-learning*), apprentissage profond (*deep-learning*) et traitement automatique de la langue et fouille du web (*natural language processing and webmining*). L'ensemble de ces matières permettent de mettre en œuvre les techniques classiques, en développant un esprit critique sur leurs limites (sur-apprentissage, grande dimension, représentativité de l'échantillon) et en utilisant des données non structurées (texte, image...). L'UE est accompagnée de séminaires professionnels sur des sujets émergents dans le champ de l'apprentissage statistique.

## Compétences ou acquis d'apprentissage à l'issue de l'UE :

Cette UE permet de maîtriser des méthodes et des outils de l'ingénieur (identification, modélisation et résolution de problèmes même non familiers et incomplètement définis, l'utilisation des approches numériques et des outils informatiques, l'analyse et la conception de systèmes) en développant l'aptitude à étudier et résoudre des problèmes complexes, à concevoir et mettre en œuvre des projets de collecte et d'analyse d'informations et à concevoir et mettre en œuvre des algorithmes prédictifs de *machine-learning*.

## Les pré-requis de l'UE :

Modélisation statistique de 2<sup>ème</sup> année, méthodes d'optimisation et d'algorithmique, aisance en R et Python.

UE 1 - Machine Learning

## APPRENTISSAGE AUTOMATIQUE

### *Machine Learning*

<i>Enseignant</i>	: François PORTIER (Ensaï) et Sébastien Da Veiga (Ensaï)
<i>Nombre d'ECTS</i>	: 2,5
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 64h
<i>Répartition des enseignements</i>	: 18h de cours et 12h d'ateliers
<i>Langue d'enseignement</i>	: Anglais (cours) et Français (ateliers)
<i>Logiciels</i>	: R
<i>Documents pédagogiques</i>	: supports de cours, bibliographie et fiches de TP
<i>Pré-requis</i>	: régression, régression logistique, convergence de lois de probabilité, théorème central-limite, algèbre linéaire ; classification par arbre ; programmation avec R ; programmation orientée objet ; optimisation: maximisation d'une fonction, algorithme de Newton

### Modalités d'évaluation :

- 1 compte-rendu de TP
- examen final

### Acquis d'apprentissage (objectifs) :

Comprendre les différents modèles de l'état de l'art (modèle linéaire: régression et classification, pénalisation, méthodes locales, SVM, forêt) ; Connaître les cas d'usage de ces modèles ; Savoir comparer empiriquement différents modèles pour une tâche donnée ; Savoir implémenter les méthodes étudiées en Python ou R

### Principales notions abordées :

Supervised learning; Regression; Classification; Empirical risk minimization; Model evaluation; Cross validation; Functional approximation; Model complexity; Large scale optimization; Stochastic gradient descent; Regularization; RIDGE and LASSO; Support Vector Machine; Kernel trick; Ensemble methods; Aggregation and Boosting; Random forest.

### Références bibliographiques :

- T. HASTIE, J. FRIEDMAN and R. TIBSHIRANI. *The elements of Statistical Learning: Data Mining, Inference and Prediction* (2nd ed.), 2009
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013), *An introduction to statistical learning*, New York: springer.
- Statistical learning with sparsity: the lasso and generalizations, T Hastie, R Tibshirani, M Wainwright – 2019  
Link: <https://web.stanford.edu/~hastie/StatLearnSparsity/>

UE 1 - Machine Learning

## APPRENTISSAGE PROFOND

### *Deep Learning*

<i>Enseignant</i>	: Benjamin GIRAULT (Ensaï)
<i>Nombre d'ECTS</i>	: 1
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 24h
<i>Répartition des enseignements</i>	: 6h de cours et 6h d'ateliers.
<i>Langue d'enseignement</i>	: Anglais
<i>Logiciels</i>	: Python (tensorflow)
<i>Documents pédagogiques</i>	: supports de cours, bibliographie et fiches de TP
<i>Pré-requis</i>	: R, Python, modélisation statistique, apprentissage statistique, optimisation de fonctions

### Modalités d'évaluation :

1 Quizz et 1 compte-rendu de TP

### Acquis d'apprentissage (objectifs) :

- identifier une tâche particulièrement adaptée pour un réseau de neurones et/ou un réseau de neurones profond
- identifier et implémenter un réseau de neurone approprié pour un tâche d'apprentissage supervisé donnée (architecture, fonction de coût, méthode d'optimisation)
- utiliser et spécialiser un réseau de neurones pré-entraîné

### Principales notions abordées :

Les réseaux de neurones profonds sont au cœur d'avancées rapides en traitement d'image et de la langue depuis les années 2010. Ce cours présente ces modèles, leur fonctionnement, ainsi que comment les utiliser.

- Principe des réseaux de neurones
- propriétés des réseaux de neurones simples
- descente de gradient
- réseaux de neurones profonds
- architectures particulières : réseaux à convolution ; réseaux récurrents.

### Références bibliographiques :

- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. 2016

UE 1 - Machine Learning

# TRAITEMENT AUTOMATIQUE DU LANGAGE ET FOUILLE DU WEB

*Natural language processing and webmining*

<i>Enseignant</i>	: Guillaume Gravier (Irisa)
<i>Nombre d'ECTS</i>	: 1,5
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 42h
<i>Répartition des enseignements</i>	: Cours : 9h – Atelier : 12h
<i>Langue d'enseignement</i>	: Anglais
<i>Logiciels</i>	: Python
<i>Documents pédagogiques</i>	: Support de cours, Supports de TP
<i>Pré-requis</i>	: Programmation avec Python, Apprentissage statistique

## Modalités d'évaluation :

Projet

## Acquis d'apprentissage (objectifs) :

- collecter des données, extraire de l'information et apparier des sources textuelles
- choisir une méthode de traitement automatique de la langue pour une tâche classique (classification, analyse de sentiment, détection > d'entités...)
- se repérer parmi le foisonnement des modèles d'étude de la langue

## Principales notions abordées :

1. What's natural language and its processing
2. The representation of words
3. The representation and classification of documents
4. Language modeling and contextual word embedding
5. Sentence-level tagging (token level tasks)
6. Sequence to sequence models and transformers
7. Overview of standard NLP tasks today

## Références bibliographiques :

- Daniel Jurafsky, James H. Martin. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*, 2nd edition, Prentice-Hall, 2009. Draft of the 3rd edition partly available at <https://web.stanford.edu/~jurafsky/slp3>.
- Yoav Goldberg. *Neural Network Methods for Natural Language Processing*. 2017. An earlier draft is freely available online at <http://u.cs.biu.ac.il/~yogo/nnlp.pdf>.
- Kevin Gimpel's lectures (Toyota Technological Institute at Chicago and UChicago) on Natural Language Processing (<https://ttic.uchicago.edu/~kgimpel/teaching/31190-s18/index.html>) and on Advanced Natural Language Processing (<https://ttic.uchicago.edu/~kgimpel/teaching/31210-s19/index.html>).

# Descriptifs des enseignements de la filière

---

UE Spécifiques filière Santé &amp; Biostatistique

## UE SPECIFIQUES FILIERE BIOSTAT

<i>Correspondant de l'UE</i>	: Myriam Vimond
<i>Nombre d'ECTS</i>	: 17
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: De 25 à 30h par ECTS
<i>Nombre d'heures d'enseignement</i>	: 279h

### Finalité des UE :

Cette filière et les UE spécifiques forment au métier de biostatisticien. Elle s'appuie sur des compléments en statistique, et fournit les outils nécessaires pour une spécialisation dans le domaine de l'expérimentation. Les cours d'épidémiologie, d'essais cliniques et l'analyse des données Omics permettent en particulier aux étudiants de recevoir une solide formation pour des applications dans le secteur de la santé.

### Structuration de l'UE :

La filière SSV inclut 4 UE spécifiques : méthodologie statistique 1&2, essais cliniques, épidémiologie/génomique.

### Compétences ou acquis d'apprentissage à l'issue de l'UE :

- Maîtriser les méthodes statistiques requises dans les trois grands domaines de :
  - l'épidémiologie (étude de la distribution dans le temps et dans l'espace des états de santé des populations humaines et l'analyse de leurs déterminants),
  - les essais cliniques (études sur les médicaments, les interventions médicales novatrices et les nouveaux matériels),
  - et l'analyse des données « Omics » (études de données génomiques, transcriptomiques, métabolomiques, protéomiques, épigénétiques et métagénomiques)
- Maîtriser la théorie sous-jacente aux méthodes statistiques (processus stochastiques, modèles de régression, statistique bayésienne, modèle de survie).
- Capacité à utiliser le raisonnement et la théorie probabiliste et statistique pour analyser des problèmes non standards survenant en médecine et en santé publique.
- Capacité à conduire des analyses médico-économiques (exemple : évaluation coût-efficacité ou coût-utilité) qui font notamment appel aux méta-analyses (savoir combiner les résultats de plusieurs essais thérapeutiques).

UE Spécifiques filière Santé &amp; Biostatistique

## UE 2 : MODELES DE REGRESSION EN BIOSTATISTIQUE

Correspondant de l'UE	: Myriam VIMOND
Nombre d'ECTS	: 5
Volume horaire de travail élève (enseignements + travail personnel)	: 146h
Nombre d'heures d'enseignement	: 69h

### Finalité de l'UE :

L'UE Modèle de régression en Biostatistique a pour objet des modèles de régression largement utilisés dans la littérature de recherche médicale. Ces modèles ont en commun de mettre en relation une variable d'intérêt avec des prédicteurs multiples : régression linéaire multiple pour une variable d'intérêt continue, régression logistique pour variable d'intérêt binaire et le modèle de risques proportionnels de Cox pour les résultats liés au temps jusqu'à l'événement. Ces modèles étudiés durant les deux premières années du cycle ingénieur à l'Ensaï sont ici approfondis et illustrés par des cas pratiques en biostatistique.

### Structuration de l'UE :

L'UE est constituée de trois parties créditées chacune de crédits ECTS selon le tableau suivant.

Plans d'expériences	18 h	2 ECTS
Modèles mixtes	24 h	1 ECTS
Analyse des Données de Survie	21 h	1,5 ECTS

*Les plans d'expériences* traitent de la conception et de l'analyse des expériences. L'objectif principal est d'obtenir un maximum d'information en un minimum d'expériences. Ils sont largement utilisés dans les domaines de l'agriculture, de la médecine, de la biologie, de la recherche marketing et de la production industrielle.

*Les modèles mixtes* (ou modèles à effets aléatoires) prennent en compte diverses structures de corrélations (intra-classe, temporelle, spatiale) en introduisant des effets aléatoires aux côtés des effets fixes habituellement considérés. Dans les biosciences, les variables de réponse peuvent être groupées et présentent une structure de corrélation particulière. Par exemple, dans le suivi d'un patient, les mesures d'un marqueur répétées chez un sujet sont corrélées et il est nécessaire d'en tenir compte dans l'analyse. Le modèle statistique doit prendre en compte la structure des corrélations.

*Les données de survie* sont définies comme un délai avant un événement (où l'événement peut être le décès, la maladie, la guérison, la rechute, ...), elles sont fréquemment rencontrées dans les études épidémiologiques ou la recherche clinique. Ces délais peuvent pour une partie des individus de l'échantillon censurés au sens où l'événement n'est pas observé. La présence de censure nécessite des techniques d'analyse de données spécifiques.

### Compétences ou acquis d'apprentissage à l'issue de l'UE :

- Maîtriser de la théorie sous-jacente à l'inférence statistique (modélisation statistique, estimation et tests d'hypothèses).
- Construire et analyser des plans d'expérience usuels.
- Conduire l'inférence statistique à l'aide d'un modèle mixte.
- Conduire l'inférence statistique pour des données de survie.

## Les pré-requis de l'UE :

Probabilité, statistique inférentielle, SAS, R (1A)

Régression linéaire, GLM, Modèle de durée, Martingales, Méthodes de rééchantillonnage (2A)

UE 2 – Modèles de Régression en Biostatistique

## PLANS D'EXPERIENCES

### *Experimental Design*

<i>Enseignant</i>	: Walter TINSSON (Université de Pau)
<i>Nombre d'ECTS</i>	: 1
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 36h
<i>Répartition des enseignements</i>	: 18h de cours
<i>Langue d'enseignement</i>	: Français
<i>Logiciels</i>	: SAS, R
<i>Documents pédagogiques</i>	: Polycopié
<i>Pré-requis</i>	: Probabilité, statistique inférentielle, SAS, R (1A) Régression linéaire (2A)

Les plans d'expérience traitent de la conception et de l'analyse des expériences. L'objectif principal est d'obtenir un maximum d'information en un minimum d'expériences. Ils sont largement utilisés dans les sciences expérimentales et les domaines l'agriculture, de la médecine, de la biologie, de la recherche marketing et de la production industrielle. Le but du cours est de donner des connaissances en plans d'expérience statistique en ce qui concerne la conception, la conduite d'expériences et l'analyse des données générées par des expériences.

*Enseignement destiné aux élèves de la filière « Data Science, Modélisation Économique & Santé » et « Data Science en Santé & Biostatistique ».*

### Principales notions abordées :

Introduction à la planification expérimentale  
Outils mathématiques pour les plans d'expérience  
Plans factoriels complets pour modèles d'ordre un  
Fractions régulières pour modèles d'ordre un  
Optimalité des plans d'expérience  
Plans d'expérience pour modèles à effets d'interactions  
Plans d'expérience pour surface de réponse

### Acquis d'apprentissage (objectifs) :

- Décrire les principes de la conception expérimentale
- Identifier les types courants et importants de conceptions expérimentales avec leurs avantages et inconvénients respectifs (puissance efficacité, coût / charge de travail),
- Choisir des modèles expérimentaux adéquats pour différents types de problèmes et de situations.
- Analyser les données générées par des plans d'expérience en utilisant l'analyse de la variance.
- Effectuer des tests post hoc.

### Modalités d'évaluation :

Examen écrit (2h), sans document.

### Références bibliographiques :

TINSSON W., *Plans d'expérience: constructions et analyses statistiques*. Springer, 2010.  
DROESBEKE, J.-J., FINE, J., SAPORTA, G. (Eds). *Plans d'expériences : Applications à l'entreprise*. Technip, 1997.  
SADO, G., SADO, M. *Les plans d'expériences: de l'expérimentation à l'assurance qualité*. Afnor, 2000.  
KHURI, A.I. et CORNELL, J.A. *Response surfaces: designs and analyses*. Routledge, 2018.  
GOUPY, J. et CREIGHTON, L. *Introduction aux plans d'expériences* Dunod, 2006

UE 2 – Modèles de Régression en Biostatistique

## MODELES MIXTES

### *Experimental Design*

<i>Enseignant</i>	: Etienne DANTAN (Université de Nantes)
<i>Nombre d'ECTS</i>	: 2
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 52h
<i>Répartition des enseignements</i>	: - Cours : 15h - Atelier : 9h
<i>Langue d'enseignement</i>	: Français
<i>Logiciels</i>	: R (nlme)
<i>Documents pédagogiques</i>	: sur Moodle
<i>Pré-requis</i>	: Probabilité, statistique inférentielle, SAS, R (1A) Régression linéaire, GLM (2A)

Dans les biosciences, les variables de réponse peuvent être groupées et présentent une structure de corrélation particulière. Par exemple, dans le suivi d'un patient, les mesures d'un marqueur répétées chez un sujet sont corrélées et il est nécessaire d'en tenir compte dans l'analyse. Le modèle statistique doit prendre en compte la structure des corrélations. Les modèles mixtes (ou modèles à effets aléatoires) prennent en compte diverses structures de corrélations (intra-classe, temporelle, spatiale) en introduisant des effets aléatoires aux côtés des effets fixes habituellement considérés. Une partie importante du cours portera sur l'utilisation et la théorie des modèles linéaires à effets mixtes (LME).

### Principales notions abordées :

Spécificités des données hiérarchiques (longitudinales).  
 Méthodes historiques d'analyses de données groupées (ANOVA pour données répétées, etc.)  
 Modèle linéaire à effets aléatoires (Définition)  
 Inférence par maximum de vraisemblance ou vraisemblance résiduelle  
 Meilleure prédiction linéaire sans biais (BLUP), Adéquation du modèle, Interprétation).  
 Prédiction des effets aléatoires.  
 Modèle linéaire généralisé mixte.

### Acquis d'apprentissage (objectifs) :

- Différencier les effets fixes et aléatoires dans un modèle mixte.
- Expliquer le principe du maximum de vraisemblance et le principe du REML.
- Implémenter un modèle mixte à l'aide de logiciel statistique (R, SAS au choix).
- Vérifier la qualité de l'ajustement du modèle mixte.
- Interpréter le résultat de l'inférence menée par modèles mixtes.
- Identifier l'usage des modèles mixtes en épidémiologie et en recherche clinique.
- Décrire l'inférence du modèle linéaire généralisé mixte.

### Modalités d'évaluation :

Examen écrit (2h).

### Références bibliographiques :

DAVIDIAN, M. et GILTINAN, D. M. *Nonlinear models for repeated measurement data*. CRC press, 1995.  
 LEE Y. , NELDER J.A., PAWITAN Y., *Generalized linear models with random effects*, CRC Press, 2018  
 PINHEIRO J.C., BATES D.M., *Mixed effects models in S and S-plus*, Springer, Berlin, 2000  
 VERBEKE G. , MOLENBERGHS G., *Linear mixed models in practice: a SAS-oriented approach*, Springer, 2012

UE 2 – Modèles de Régression en Biostatistique

## ANALYSE DE DONNEES DE SURVIE

### *Survival Data Analysis*

<i>Enseignant</i>	: Lisa BELIN (Sorbonne Université- INSERM- AP-HP), David HAJAGE (Sorbonne Université- INSERM- AP-HP)
<i>Nombre d'ECTS</i>	: 2
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 52h
<i>Répartition des enseignements</i>	: - Cours : 18h - Atelier : 6h
<i>Langue d'enseignement</i>	: Français
<i>Logiciels</i>	: R (survival)
<i>Documents pédagogiques</i>	: Polycopié et autres documents sur Moodle
<i>Pré-requis</i>	: Probabilité, statistique inférentielle, SAS, R (1A) Régression linéaire, GLM, Modèle de durée, Martingales, Méthodes de rééchantillonnage (2A)

Le cours sera illustré de nombreux exemples et des exercices seront, à chaque étape, proposés aux étudiants pour vérifier qu'ils ont bien compris et acquis l'essentiel des méthodes. Ces exemples et exercices seront principalement issus de travaux effectués en recherche clinique et en épidémiologie. Cependant, le caractère général de ces méthodes, utiles dans bien d'autres domaines d'application, sera clairement établi.

### Principales notions abordées :

Particularités des données de survie (censure, troncature...)

Fonctions de risque et de survie, et autres représentations

Estimation non-paramétrique des fonctions de survies - Comparaisons non-paramétriques des distributions de survies

Modèles semi-paramétriques via le modèle de Cox - Cas des variables dont la valeur dépend du temps.

Vérification des hypothèses du modèle de Cox : hypothèse des hasards proportionnels et hypothèse de log-linéarité

Modèles avancés en survie : introduction de la notion de risque compétitif

Utilisation des scores de propension dans un modèle de Cox

### Acquis d'apprentissage (objectifs) :

- Définir la relation entre la fonction de survie, la fonction de répartition, la fonction de risque instantané et la densité.
- Identifier la présence de censure et de troncature dans un jeu de données
- Décrire et interpréter des données de survie à l'aide de l'estimateur de Kaplan-Meier.
- Formuler des questions de recherche impliquant des données de survie en tant que problèmes de régression.
- Résoudre les problèmes classiques de comparaison de plusieurs échantillons des distributions de survie.
- Implémenter un modèle de durée à l'aide de logiciel statistique (R)
- Vérifier la qualité de l'ajustement du modèle pour des données de survies.
- Interpréter les résultats de l'inférence sur des données de survies via de modèle semiparamétrique multivarié
- Intégrer des covariables dépendantes du temps dans le modèle à risques proportionnels
- Identifier des écarts aux principales hypothèses des modèles

### Modalités d'évaluation :

Un examen écrit (2h). Un devoir maison à mi-parcours du volume horaire consistera à une mise en pratique des notions étudiées et modèlera la note finale de l'étudiant en fonction de la qualité du devoir.

## Références bibliographiques :

- C. HILL, C. COM-NOUGUE, A. KRAMAR, T. MOREAU, J. O'QUIGLEY, R. SENOUSI, C. CHASTANG, *Analyse statistique des données de survie* (2e éd.), Collection statistique en biologie et médecine, Flammarion, 1996.
- A. KRAMAR et S. MATHOULIN-PELISSIER, *Méthodes biostatistiques appliquées à la recherche clinique en cancérologie*. Collection : *L'innovation thérapeutique en cancérologie*, Éd JOHN LIBBEY EUROTEXT, 2011
- H. VAN HOUWELINGEN, H. PUTTER. *Dynamic prediction in clinical survival analysis*. CRC Press, 2011. JP. KLEIN, HC VAN HOUWELINGEN, JG. IBRAHIM (ed.). *Handbook of survival analysis*. CRC Press, 2016.

UE Spécifiques filière Santé &amp; Biostatistique

## UE 3 : MODELISATION STOCHASTIQUE ET STATISTIQUE

Correspondant de l'UE	: Myriam VIMOND
Nombre d'ECTS	: 5
Volume horaire de travail élève (enseignements + travail personnel)	: 139h
Nombre d'heures d'enseignement	: 72h

### Finalité de l'UE :

L'UE Modélisation stochastique et statistique a pour objet l'étude de modèles fréquemment utilisées en Biostatistique. Elle couvre l'étude des processus stochastiques couramment utilisés en épidémiologie ou en génomique, la modélisation bayésienne applicable à des problèmes complexe d'analyse des données, notamment en méta-analyse, ainsi que la considération du traitement des données manquantes spécifique aux essais cliniques.

### Structuration de l'UE :

L'UE est constituée de trois parties créditées chacune de crédits ECTS selon le tableau suivant.

Statistique des Processus	21 h	2 ECTS
Statistique Bayésienne	18 h	1 ECTS
Méta-Analyse	18 h	1 ECTS
Traitement des données manquantes	15 h	1 ECTS

*La Statistique des Processus* traitent des différents modèles définis à partir de processus stochastique. L'objectif principal est *deriver* l'inférence statistique à partir de modèle simple faisant intervenir les processus de Poisson, les chaînes de Markov, les modèles de Markov cachées et les chaîne de Markov à sauts.

*La Statistique Bayésienne* sont dorénavant applicables à des problèmes complexes d'analyse de données. Elles sont notamment particulièrement adaptées pour mener l'inférence statistique de modèles à variables latentes (i.e., modèles mixtes, modèles de mélange etc...), leur comparaison et la prédiction de données dans ce contexte ainsi que pour la prise en compte de sources d'incertitude multiples en apprentissage statistique.

*Une Méta-Analyse* est une analyse statistique qui a pour but de combiner les résultats de plusieurs essais thérapeutiques, examiner les sources d'hétérogénéité et les biais potentiels. Elle repose généralement sur des modèles mixtes ou sur des modèles bayésiens.

*Le traitement des données manquantes* est souvent à considérer en recherche clinique. Si elles sont négligées ou mal gérées lors des procédures analytiques, elles peuvent biaiser ou réduire la puissance de l'étude ou conduire à des conclusions invalides.

### Compétences ou acquis d'apprentissage à l'issue de l'UE :

- Maîtriser de la théorie sous-jacente à l'inférence statistique fréquentiste (modélisation statistique, estimation et tests d'hypothèses) et bayésienne (loi a posteriori, intervalles de crédibilité, ...).
- Mettre en œuvre une modélisation utilisant un processus stochastique à des fins de modélisation statistique.
- Traiter des problèmes courants d'analyse de données avec des méthodes statistiques bayésiennes.

- Mettre en application la ou les méthodes de traitements des données manquantes adaptées à un problème donné.
- Planifier une méta-analyse dans une revue systématique dans le domaine des essais cliniques.
- Conduire une méta-analyse standard et en réseaux.

### Les pré-requis de l'UE :

Probabilité, statistique inférentielle, SAS, R (1A)

Régression linéaire, GLM, Modèle de durée, Chaînes de Markov, Calcul Bayésien, Méthodes de rééchantillonnage (2A)

UE 3 – Modélisation Statistique et Stochastique

## STATISTIQUE DES PROCESSUS

*Statistic of Stochastic Processes*

<i>Enseignant</i>	: Lionel TRUQUET (ENSAI)
<i>Nombre d'ECTS</i>	: 2
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 46h
<i>Répartition des enseignements</i>	: - Cours : 18h - Atelier : 3h
<i>Langue d'enseignement</i>	: Français
<i>Logiciels</i>	: R (coda, bps, spline, HiddenMarkov, depmixS4)
<i>Documents pédagogiques</i>	: Polycopié
<i>Pré-requis</i>	: Probabilité, statistique inférentielle, SAS, R (1A) Chaîne de Markov, Calcul Bayésien, Martingales, Méthodes de rééchantillonnage (2A))

Un processus stochastique est un ensemble de variables aléatoires indexées par le temps ou l'espace. La modélisation stochastique est utilisée pour la modélisation de phénomènes temporels ou spatiaux dans de nombreux domaines comme la biologie, la fiabilité ou la finance. Ce cours a pour objectif d'étudier les propriétés théoriques de différents types de processus stochastiques, notamment les chaînes de Markov, les processus de Poisson et les processus de Markov à sauts. Ce cours sera également illustré par des des exemples de modélisation et d'inférence statistique.

*Enseignement destiné aux élèves de la filière « Data Science, Modélisation économique & Santé » et « Data Science en Santé & Biostatistique ».*

### Principales notions abordées :

Processus Stochastiques - Inférence paramétrique  
 Processus de Renouvellement  
 Processus de Poisson - Points de rupture  
 Chaîne de Markov - Etats absorbants - Modèle de Markov Caché  
 Processus de Markov à sauts – Modèles phylogénétiques  
 Files d'attente (*uniquement pour « Génie Statistique »*)  
 Aperçu des chaînes de Markov inhomogène (*uniquement pour « Génie Statistique »*)

### Acquis d'apprentissage (objectifs) :

- Définir les processus stochastiques suivants chaîne de Markov, processus de Poisson, processus de Markov à sauts)
- Effectuer des calculs de base pour les processus de Poisson et les chaînes de Markov, y compris les modèles de files d'attente simples.
- Décrire des modèles statistiques faisant intervenir des processus stochastiques
- Développer les méthodes d'inférence fréquentistes ou bayésiennes dérivées des modèles de processus stochastiques.

### Modalités d'évaluation :

Examen écrit

### Références bibliographiques :

ASMUSSEN, S. Applied probability and queues. Second edition. Springer 2003.  
 BOSQ, D. Statistique mathématique et statistique des processus. Lavoisier, 2012.

DELMAS, J-F., JOURDAIN, B. Modèles aléatoires. Applications aux sciences de l'ingénieur et du vivant. Springer 2006.  
FOATA, D., FUCHS, A. Processus Stochastiques (2<sup>e</sup> éd.). Dunod 2004.  
PARDOUX, E. Processus de Markov et applications. Algorithmes, génome et finance. Dunod 2007.

UE 3 – Modélisation Statistique et Stochastique

## TRAITEMENT DE LA DONNÉE MANQUANTE

### *Handling Missing Data*

<i>Enseignant</i>	: Mélanie PRAGUE (INRIA)
<i>Nombre d'ECTS</i>	: 1
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 30h
<i>Répartition des enseignements</i>	: - Cours : 6h - Atelier : 9h
<i>Langue d'enseignement</i>	: Français, support écrit en anglais
<i>Logiciels</i>	: R
<i>Documents pédagogiques</i>	: Notes de Cours sur Moodle
<i>Pré-requis</i>	: Probabilité, statistique inférentielle, R (1A) Chaîne de Markov, Calcul Bayésien, Méthodes de rééchantillonnage (2A) Modèles mixtes (3A)

Les données manquantes sont fréquemment rencontrées dans la recherche clinique. Si elles sont négligées ou mal gérées lors des procédures analytiques, elles peuvent biaiser réduire la puissance de l'étude ou conduire à des conclusions invalides. Le cours présente les concepts clés concernant les données manquantes dans la recherche clinique, décrit les mécanismes de données manquantes, ainsi que les méthodes spécifiques de traiter des données incomplètes. En outre le cours aborde les notions d'analyse des cas complets, le modèle à effets mixtes pour les mesures répétées, les méthodes de pondération, l'imputation simple, l'imputation multiple, analyse de sensibilité. Ce cours comprend deux ateliers sous R mettant en application les différentes techniques du cours.

### Principales notions abordées :

Typologie des données manquantes : MCAR, MAR and MNAR  
 Utilisation de DAG pour comprendre les données manquantes  
 Méthodes Ad-hoc: Analyse des cas complet (CCA) / LOCF / Imputation simple  
 Hypothèse MAR: Approches par maximum de vraisemblance  
 Inverse Probability Weighting (IPW)  
 Imputation multiple: loi re Rubin, equation chaînée, algorithmes MCMC  
 Modèles de données manquantes non monotones: algorithmes MCMC  
 Hypothèse MNAR : Selection de modèle et Pattern-Mixture-Models (PMM)  
 Analyse de sensibilité sous l'hypothèse MNAR

### Acquis d'apprentissage (objectifs) :

- Différencier les trois mécanismes de données manquantes et décrire leur incidence sur les résultats d'une étude.
- Décrire les principales techniques statistiques pour traiter les données manquantes dans les ensembles de données épidémiologiques et cliniques.
- Identifier les forces et les limites des méthodes les plus fréquemment utilisées pour traiter les données manquantes dans diverses situations de données manquantes.
- Traiter les données manquantes dans une base de données.
- Appliquer et interpréter une méthode d'imputation multiple à l'aide de R.

### Modalités d'évaluation :

Evaluation à rendre à l'issue de l'atelier

## Références bibliographiques :

LITTLE, Roderick JA et RUBIN, Donald B. *Statistical analysis with missing data*. John Wiley & Sons, 2019.

VAN BUUREN, Stef. *Flexible imputation of missing data*. CRC press, 2018.

MOLENBERGHS, Geert et KENWARD, Michael. *Missing data in clinical studies*. John Wiley & Sons, 2007.

UE 3 – Modélisation Statistique et Stochastique

## STATISTIQUE BAYESIENNE

### *Bayesian Inference*

<i>Enseignant</i>	: Sophie ANCELET (IRSN), Eric PARENT (AgroParisTech/INRA)
<i>Nombre d'ECTS</i>	: 1
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 33h
<i>Répartition des enseignements</i>	: - Cours : 9h - Atelier : 9h
<i>Langue d'enseignement</i>	: Français
<i>Logiciels</i>	: R (packages rjags, rstan, coda, ggmmcmc), JAGS, STAN
<i>Documents pédagogiques</i>	: Slides + Documents R Markdown pour corrections TP
<i>Pré-requis</i>	: Probabilité, statistique inférentielle, SAS, R (1A) Régression linéaire, GLM, Chaîne de Markov, Calcul bayésien (2A)

La statistique bayésienne se distingue par l'utilisation d'une loi de probabilité sur l'ensemble des objets (ex : paramètres, modèles, etc...) sur lesquels l'inférence est menée. Etant donné un modèle probabiliste paramétrique (cas exclusivement considéré dans ce cours), les paramètres inconnus du modèle sont traités comme des variables aléatoires : une loi de probabilité dite loi a priori leur est assignée afin de décrire un état de connaissances (i.e., incertitude épistémique) avant la collecte de données. La formule de Bayes permet ensuite une mise à jour de cet état de connaissances via l'assimilation de données observées afin d'obtenir une loi de probabilité dite loi a posteriori sur les paramètres inconnus. En pratique, la loi a posteriori n'a généralement pas une forme analytique connue. Différentes approximations de la loi a posteriori seront présentées, notamment les Algorithmes Monte-Carlo par Chaînes de Markov (MCMC) (formes classiques et plus avancées) dans la continuité du cours de calcul bayésien de 2A. Une initiation aux logiciels gratuits JAGS et STAN (packages R « rjags » et « rstan ») sera proposée lors des ateliers. Le cours traitera enfin du problème de la sélection et de la validation prédictive de modèles sous le paradigme bayésien, et sera illustrée par des exemples empreints aux sciences de la vie au sein desquelles l'emploi des méthodes statistiques bayésiennes progresse considérablement.

*Enseignement destiné aux élèves de la filière « Data Science & Modélisation », « Data Science en Santé » et « Data Science & Economie ».*

### Principales notions abordées :

Formule de Bayes / Loi a priori / Loi a posteriori / Loi prédictive a posteriori / Incertitude épistémique  
 Estimateurs bayésiens / Intervalles de crédibilité  
 Lois a priori conjuguées / Lois a priori de Jeffreys  
 Modèles à variables latentes et représentation hiérarchique / Graphe acyclique orienté  
 Approximations déterministes de la loi a posteriori  
 Algorithmes MCMC (Principe, Gibbs/Metropolis-Hastings/Dynamiques Hamiltoniennes, diagnostics de convergence)  
 Validation prédictive bayésienne (posterior predictive check, cross validation)  
 Sélection bayésienne de modèles (Facteur de Bayes, Deviance Information Criterion, Widely Applicable Information Criterion)

### Acquis d'apprentissage (objectifs) :

- Expliquer le principe général de la démarche statistique bayésienne
- Faire un choix pertinent de lois a priori
- Formuler l'écriture probabiliste d'un modèle à variables latentes sous forme hiérarchique et donner sa représentation sous forme de graphe acyclique orienté
- Calcul de lois a posteriori (*uniquement pour SSV*)

- Mener l'inférence bayésienne de modèles classiques (e.g., régression linéaire, GLM) et de modèles à variables latentes (e.g., modèles mixtes) à l'aide des packages R « rjags » et « rstan »
- Mener une étude de convergence d'un algorithme MCMC
- Comparer différents modèles à l'aide de critères de sélection bayésiens
- Valider un modèle – du point de vue prédictif - sous le paradigme bayésien

## Modalités d'évaluation :

Devoir maison

## Références bibliographiques :

Collectif BIOBAYES: ALBERT I., ANCELET S., DAVID O., DENIS J.B., MAKOWSKI D., PARENT E., SOUBEYRAND S. *Méthodes statistiques bayésiennes. Bases théoriques et applications en alimentation, environnement et génétique. Ellipses, 2015*

GELMAN, A., CARLIN, J. B., STERN, H. S. and. RUBIN, D. B. *Bayesian data analysis*. CRC Press, 2013

## UE 3 – Modélisation Statistique et Stochastique

## META-ANALYSE

### Meta-Analysis

<i>Enseignant</i>	: Pauline LE NOUVEAU
<i>Nombre d'ECTS</i>	: 1
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 30h
<i>Répartition des enseignements</i>	: 12h de cours, 6h d'atelier
<i>Langue d'enseignement</i>	: Anglais
<i>Logiciels</i>	: R, SAS et WinBugs (ou Jags)
<i>Documents pédagogiques</i>	: sur Moodle
<i>Pré-requis</i>	: Probabilité, statistique inférentielle, SAS, R (1A) Régression, GLM, Modèle de durée, Chaîne de Markov, Calcul Bayésien (2A) Statistique Bayésienne, Modèles Mixtes, Analyse de Données de Survie (3A)

De nombreuses expériences sont réalisées chaque année pour analyser les effets d'un traitement sur la santé ou de pratiques agricoles sur la production et l'environnement. Une méta-analyse est une analyse statistique qui a pour but de combiner les résultats de plusieurs essais thérapeutiques, examiner les sources d'hétérogénéité et les biais potentiels. Le cours couvrira la théorie et l'utilisation des méthodes méta-analytiques les plus courantes, l'interprétation et les limites des résultats de ces méthodes, les procédures de diagnostic et certains sujets avancés axés sur les applications de santé publique.

*Enseignement destiné aux élèves de la filière « Data Science en Santé & Biostatistique » et « Data Science, Modélisation économique & Santé ».*

### Principales notions abordées :

Utilisation de la méta-analyse dans les essais cliniques / évaluation économique de la santé  
 Méta-analyse versus essais cliniques randomisés  
 Planification d'une méta-analyse dans une revue systématique  
 Estimation de la différence de traitement pour des données binaires  
 Estimation de la différence de traitement pour des données normalement distribuées  
 Estimation de la différence de traitement pour des données ordinales  
 Estimation de la différence de traitement pour des données de survie  
 Combinaison d'estimations de la différence de traitement (modèles à effets fixes et modèles mixtes)  
 Hétérogénéité et méta-régression  
 Biais de sélection / publication  
 Recommandations pour produire une Méta-analyse

### Acquis d'apprentissage (objectifs) :

- Expliquer le rôle des revues systématiques et des méta-analyses en santé;
- Décrire différentes façons de mesurer l'hétérogénéité entre les études;
- Mener une revue systématique de la littérature afin de sélectionner les études dans la méta-analyse;
- Conduire une méta-analyse standard et en réseaux avec estimation de la taille de l'effet et interprétation des résultats.

### Modalités d'évaluation : Dossier écrit (10h)

## Références bibliographiques :

- WHITEHEAD, A. *Meta-analysis of controlled clinical trials*. Vol. 7. John Wiley, 2002
- SCHWARZER, G., CARPENTER, J. R., et RÜCKER, G.. *Meta-analysis with R*. Cham : springer, 2015.
- HIGGINS, J., THOMAS, J., CHANDLER, J., et al. (ed.). *Cochrane handbook for systematic reviews of interventions*. Wiley, 2019.
- DIAS, S., WELTON, N.J., SUTTON, A.J., et al. *NICE DSU technical support document 2: a generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials*. 2011.
- WELTON, N. J., SUTTON, A. J., COOPER, N., ABRAMS, K. R., & ADES, A. E.. *Evidence synthesis for decision making in healthcare* (Vol. 132). Wiley, 2012
- SUTTON, A. J., ABRAMS, K. R., JONES, D. R., JONES, D. R., SHELDON, T. A., & SONG, F.. *Methods for meta-analysis in medical research* (Vol. 348). Chichester: Wiley, 2000

UE Spécifiques filière Santé &amp; Biostatistique

## UE 4 : ESSAIS CLINIQUES

<i>Correspondant de l'UE</i>	: Myriam VIMOND
<i>Nombre d'ECTS</i>	: 5
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 160h
<i>Nombre d'heures d'enseignement</i>	: 75h

### Finalité de l'UE :

Le développement d'une molécule est un processus complexe qui dure entre 5 et 10 ans. Il comporte de nombreux jalons afin de montrer aux autorités que la molécule est efficace tout en étant suffisamment peu toxique pour l'homme. Les essais cliniques sont des étapes essentielles de ce développement et ont pour but l'évaluation de nouveaux traitements ou de nouvelles modalités de soins. L'UE essai clinique a pour but d'initier les étudiants à une lecture plus critique de la littérature sur les essais cliniques et fournira les outils pour concevoir leur propre protocole.

### Structuration de l'UE :

L'UE est constituée de quatre parties créditées chacune de crédits ECTS.

Essais cliniques	42 h	3 ECTS
Echelle de mesure en santé	15 h	1 ECTS
Pharmacométrie	18 h	1 ECTS

Le cours Essais cliniques a pour objet de familiariser les étudiants à la nature et la structure des données recueillies selon le type d'études (sur les médicaments, les interventions médicales novatrices et les nouveaux matériels ...), le déroulement et les enjeux de telles études, ainsi que sur les méthodes statistiques adaptées pour traiter ces données.

Le Projet Essais cliniques suivra les étapes majeures du déroulement d'une étude clinique, notamment la réflexion sur le design et le type d'analyse statistique, l'évaluation des caractéristiques opérationnelles à l'aide d'un plan de simulation, la présentation du synopsis, l'écriture du protocole et du plan d'analyse statistique de l'étude.

Le cours Echelle de mesures en santé introduit l'évaluation de la qualité de vie des patients qui est un des enjeux dans le domaine de l'évaluation des stratégies thérapeutiques des maladies chroniques, le maintien ou l'amélioration de la qualité de vie des patients tout au long de l'évolution de la maladie. La qualité de vie n'étant pas directement observable, son évaluation fait appel à des méthodes et concepts développés en psychométrie.

La Pharmacométrie consiste à modéliser les données obtenues lors des essais cliniques sur des médicaments, afin de mieux comprendre et prédire l'efficacité et la toxicité de ces produits ainsi que les sources de variabilité chez les patients. Les modèles pharmacocinétiques/pharmacodynamiques (PK/PD) sont souvent non linéaires, reflétant les phénomènes de saturation ou les chaînes complexes de réactions biochimiques.

### Compétences ou acquis d'apprentissage à l'issue de l'UE :

- Expliquer l'intérêt et les hypothèses sous-jacentes aux différentes méthodes utilisées lors de la planification, la conduite et l'analyse statistique des essais cliniques.
- Citer les grandes lignes de la réglementation et de l'éthique qui encadre les essais cliniques en France.

- Évaluer de manière critique les facteurs qui contribuent à une conception d'essai appropriée pour une question de recherche, y compris la sélection des paramètres appropriés, le choix de la taille de l'échantillon, l'analyse des données, la présentation des résultats.
- Contribuer à la conception, à la rédaction et à l'évaluation d'un protocole d'essais cliniques.
- Expliquer les enjeux de la validation d'un questionnaire pour produire une mesure d'un concept
- Décrire les grandes théories psychométriques (théorie classique, théorie de mesure de Rasch, théorie de mesure à l'item...), ainsi que les différences de traitement statistique des données en fonction de la théorie utilisée.
- Développer et évaluer un modèle pour l'évolution temporelle de la concentration et de la réponse du médicament dans diverses situations.
- Conduire l'inférence pour les modèles non-linéaire à effets mixtes en utilisant les logiciels dédiés.

### Les pré-requis de l'UE :

Probabilité, statistique inférentielle, SAS, R (1A)

Régression linéaire, GLM, Modèle de durée, Chaîne de Markov, Calcul Bayésien, Méthodes de rééchantillonnage (2A)

Modèles mixtes, Plan d'expérience, Analyse des Données de Survies (3A)

UE 4 – Essais Cliniques

## ESSAIS CLINIQUES

### *Clinical Trials*

<i>Enseignant</i>	: David HAJAGE (Sorbonne Université - INSERM - AP-HP - Enseignant coordinateur), Lisa BELIN (Sorbonne Université - INSERM - AP-HP), Yann DE RYCKE (Sorbonne Université - INSERM - AP-HP) et Caroline PETIT (Sanofi)
<i>Nombre d'ECTS</i>	: 3
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 84h
<i>Répartition des enseignements</i>	: - Cours : 18h - Projet : 24h
<i>Langue d'enseignement</i>	: Français
<i>Logiciels</i>	: R, SAS
<i>Documents pédagogiques</i>	: Polycopié
<i>Pré-requis</i>	: Probabilité, statistique inférentielle, SAS, R (1A) Régression linéaire, GLM, Modèle de durée, Calcul Bayésien, Méthodes de rééchantillonnage (2A) Modèles mixtes, Analyse des données de survie, Plans d'expériences (3A)

Le développement d'une molécule est un processus complexe qui dure entre 5 et 10 ans. Il comporte de nombreux jalons afin de montrer aux autorités que la molécule est efficace tout en étant suffisamment peu toxique pour l'homme. Les essais cliniques sont des étapes essentielles de ce développement.

La nature et la structure des données recueillies dans le cadre d'essais cliniques (qui incluent les études sur les médicaments, mais aussi sur les dispositifs médicaux ou les stratégies de soin de manière générale) nécessitent de recourir à des plans expérimentaux et des méthodes statistiques adaptées.

La première partie du cours présentera le plan expérimental le plus utilisés pour les essais cliniques de phase III (l'essai contrôlé randomisé) ainsi que le déroulement d'une étude du point de vue du biostatisticien. Les suivantes s'attacheront à détailler certains plans expérimentaux et méthodes d'analyse répondant à des objectifs particuliers (essais précoces, essais d'équivalence, analyses intermédiaires).

Dans un deuxième temps, les étudiants travailleront sur un projet qui est une mise en situation de conception d'un essai clinique dans le développement d'une molécule pour une pathologie précise. Ce projet demande un important travail de programmation, de réflexion statistique, un effort de recherche sur la pathologie étudiée ainsi qu'une bonne compréhension des recommandations internationales sur l'analyse statistique et la rédaction d'un rapport dans le cadre d'un essai clinique.

### Principales notions abordées :

Présentation générale des essais cliniques  
 Prise en compte des biais et choix d'un plan expérimental adapté aux objectifs de l'essai clinique  
 Les différentes étapes d'un essai clinique  
 Protocole d'essai clinique, plan d'analyse statistique, rapport d'essai clinique  
 Acculturation aux aspects réglementaires et éthiques  
 Puissance et nombre de sujets nécessaires  
 Essais d'équivalence et de non infériorité  
 Essais de phase I et essais de phase II / Analyses intermédiaires

### Acquis d'apprentissage (objectifs) :

- Expliquer les enjeux et les différents plans expérimentaux et méthodes statistiques utilisées dans un essai clinique
- Décrire les biais possibles dans l'évaluation thérapeutiques et les mesures mises en place pour les éviter

- Calculer un nombre de sujets nécessaires tenant compte de l'objectif, du critère d'évaluation et du plan expérimental de l'essai clinique
- Évaluer de manière critique les facteurs qui contribuent à une conception d'essai appropriée pour une question de recherche, y compris la sélection des paramètres appropriés, le choix de la taille de l'échantillon, l'analyse des données, la présentation des résultats.
- Contribuer à la conception, à la rédaction et à l'évaluation d'un protocole d'essais cliniques.
- Construire un plan de simulation et l'appliquer afin de répondre à une problématique liée à l'essai.

## Modalités d'évaluation :

Examen écrit pour la partie cours (coefficient 1) : questions rédactionnelles sur un article rapportant les résultats d'un essai clinique. Documents autorisés.

Le projet (coefficient 2) est réalisé en général par groupe de 2 ou 3 élèves. Chaque équipe projet suivra les étapes majeures du déroulement d'une étude clinique, notamment la réflexion sur le design et le type d'analyse statistique, l'évaluation des caractéristiques opérationnelles à l'aide d'un plan de simulation (en R), la présentation du synopsis, l'écriture du protocole et du plan d'analyse statistique de l'étude. Pour chaque projet, le contexte sera fourni sous la forme d'une publication scientifique avec le protocole associé. Chaque groupe devra alors comprendre, évaluer et construire l'étape suivante du développement de la molécule associée et proposer un essai clinique allant dans cette direction. Les étudiants sont évalués sous deux formes. Au cours du projet, les élèves seront amenés à présenter l'évolution de leur travail sous un format de sessions courtes et synthétiques à l'ensemble de la classe. Une note de participation sera attribuée. A la fin du projet, les élèves doivent remettre un rapport statistique d'une vingtaine de pages donnant lieu à une soutenance réalisée devant des professionnels des essais cliniques. Le rapport ainsi que la soutenance seront réalisés en anglais, langue de travail utilisée dans l'industrie pharmaceutique.

## Références bibliographiques :

COOK, T. D., DEMETS, D. L. *Introduction to Statistical Methods for Clinical Trials*. CRC Press, 2007

FRIEDMAN, L. M., FURBERG, C. D., DEMETS, D. L., et al. *Fundamentals of clinical trials*. Springer, 2015.

CLEOPHAS, T. J. et ZWINDERMAN, A. H. *Statistics applied to clinical studies*. Springer, 2012.

EVANS, Scott et TING, Naitee. *Fundamental concepts for new clinical trialists*. CRC Press, 2015.

FRIEDMAN, Lawrence M., FURBERG, Curt D., DEMETS, David L., et al. *Fundamentals of clinical trials*. Springer, 2015.

CHEN, Ding-Geng Din, PEACE, Karl E., et ZHANG, Pinggao. *Clinical trial data analysis using R and SAS*. CRC Press, 2017.

DMITRIENKO, Alex et PULKSTENIS, Erik (ed.). *Clinical trial optimization using R*. CRC Press, 2017.

UE 4 – Essais Cliniques

## PHARMACOMETRIE

### Pharmacometrics

<i>Enseignant</i>	: Jérémie GUEDJ (INSERM, responsable du cours) Alexandra LAVALLEY-MORELLE (INSERM)
<i>Nombre d'ECTS</i>	: 1
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 48h
<i>Répartition des enseignements</i>	: - Cours : 6h - Atelier : 12h
<i>Langue d'enseignement</i>	: Français
<i>Logiciels</i>	: Monolix R
<i>Documents pédagogiques</i>	: sur Moodle
<i>Pré-requis</i>	: Probabilité, statistique inférentielle, Optimisation, R (1A) Régression linéaire, GLM, Calcul Bayésien (2A) Modèles mixtes (3A)

La pharmacométrie consiste à modéliser les données obtenues lors des essais cliniques sur des médicaments, afin de mieux comprendre et prédire l'efficacité et la toxicité de ces produits ainsi que les sources de variabilité chez les patients. Les modèles pharmacocinétiques (PK) décrivent la relation entre la dose de produit administrée et sa concentration dans l'organisme, et les modèles pharmacodynamiques (PD) décrivent la relation entre la concentration et son effet. Ces modèles pharmacocinétiques/pharmacodynamiques (PK/PD) sont souvent non linéaires, reflétant les phénomènes de saturation ou les chaînes complexes de réactions biochimiques. La variabilité biologique entre des individus ou des groupes d'individus dans une population nécessite d'utiliser des modèles hiérarchiques appelés modèles non-linéaires à effets mixtes.

### Principales notions abordées :

Introduction au processus de développement de médicaments

Principes des relations PK/PD, modèles pharmacocinétiques/pharmacodynamiques (PK/PD)

Modèles de régression non linéaires (écriture sous forme d'équation différentielle)

Modèles de régression non linéaires à effets mixtes (Inférence par maximum de vraisemblance, Algorithme EM, SAEM)

Construction et évaluation de modèles / Monolix

Optimisation de protocoles dans les modèles non linéaires simples ou mixtes, théorie et applications, simulation d'essais cliniques

### Acquis d'apprentissage (objectifs) :

- Identifier la classe appropriée de modèles PK / PD pour décrire les différentes relations PK / PD.
- Décrire les modèles de régression non linéaires à effets mixtes.
- Expliquer les principes des différentes méthodes d'inférence pour les modèles non linéaires à effets mixtes.
- Proposer un modèle pour l'évolution temporelle de la concentration et de la réponse du médicament dans diverses situations
- Utiliser des logiciels type Monolix et R en pharmacométrie.
- Interpréter les résultats issus des analyses en pharmacométrie.
- Expliquer et communiquer les résultats d'une analyse pharmacométrique.

### Modalités d'évaluation :

L'évaluation se fera sur projet, comprenant au minimum une analyse de données reprenant les concepts expliqués en TP (travail évalué à deux jours pleins).

### Références bibliographiques :

GABRIELSSON, J. *PKPD Data Analysis: Concepts and Applications (5th Edition)*. Swedish Pharmaceutical Press, 2015.  
LAVIELLE, M. *Mixed effects models for the population approach: models, tasks, methods and tools*. CRC Press, 2014.  
ETTE, E. et WILLIAMS, P. (ed.). *Pharmacometrics: the science of quantitative pharmacology*. Wiley, 2013.  
BAZZOLI, C., BERTRAND, J., et COMETS, E. *La pharmacométrie*. Techniques de l'Ingenieur, 2014.

UE 3 – Essais Cliniques

## ECHELLE DE MESURE EN SANTE

### *Scale Measurement in Health*

<i>Enseignant</i>	: Jean Benoît HARDOUIN (Université de Nantes)
<i>Nombre d'ECTS</i>	: 1
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 30h
<i>Répartition des enseignements</i>	: - Cours : 12h - Atelier : 3h
<i>Langue d'enseignement</i>	: Français
<i>Logiciels</i>	: STATA
<i>Documents pédagogiques</i>	: sur Moodle
<i>Pré-requis</i>	: Probabilité, statistique inférentielle, Statistique exploratoire multivariée (1A) Régression linéaire, GLM (2A) Modèle Mixte (3A, pas indispensable)

Dans le domaine de l'évaluation des stratégies thérapeutiques des maladies chroniques, le maintien ou l'amélioration de la qualité de vie des patients tout au long de l'évolution de la maladie est devenu un des enjeux au cours de ses dernières années. La qualité de vie n'étant pas directement observable, son évaluation fait appel à des méthodes et concepts développés en psychométrie. Différentes théories de psychométrie coexistent en se différenciant sur des points particuliers : soit obtenir une mesure simple du concept que l'on cherche à mesurer (psychométrie classique), soit aboutir à une mesure ayant de bonnes propriétés métrologiques (théorie de la mesure de Rasch), soit rechercher un modèle statistique le plus adéquat possible (théorie de réponse à l'item).

### Principales notions abordées :

Construction d'un questionnaire

Les grandes étapes de la validation d'un questionnaire (validité, fiabilité, sensibilité au changement)

Les grandes théories psychométriques : points communs et spécificités

La théorie classique des tests (CTT)

L'extension de la CTT à l'aide des modèles d'équations structurelles (SEM)

La théorie de mesure de Rasch (RMT)

La théorie de réponse à l'item (IRT)

Méthodes de validation adaptée à chaque théorie psychométrique

L'analyse de données issues de questionnaires

Atelier pratique : la validation d'un questionnaire en CTT et RMT : application sur le questionnaire "Impact of cancer"

### Acquis d'apprentissage (objectifs) :

- Expliquer les enjeux de la validation d'un questionnaire pour produire une mesure d'un concept
  - Mettre en œuvre la construction d'une échelle de mesure d'un concept
  - Identifier les différentes étapes d'une validation de questionnaire
- Présenter les grandes théories psychométriques (théorie classique, théorie de mesure de Rasch, théorie de mesure à l'item...) et les principes des méthodes statistiques permettant de valider un questionnaire
- Analyser des données issues d'un questionnaire psychométrique
  - Choisir le traitement statistique approprié en fonction des caractéristiques des données et des objectifs de l'étude
  - Interpréter les mesures obtenues dans chaque théorie psychométrique

Modalités d'évaluation : Examen écrit.

## Références bibliographiques :

B FALISSARD, Mesurer la subjectivité en santé (2e éd.) Collection évaluation et statistique, Masson, 2008

HCW de Vet, CB Terwee, LB Mokkink, DL Knol. Measurement in Medicine: A Practical Guide (Practical Guides to Biostatistics and Epidemiology) Cambridge University Press, 2011.

PM FAYERS, D MACHIN, Quality of Life : assessment, analysis and interpretation (2nd ed.), Wiley, 2007

UE Spécifiques filière Santé &amp; Biostatistique

## UE 5 : EPIDEMIOLOGIE & DONNEES OMIQUES

Correspondant de l'UE	: Myriam VIMOND
Nombre d'ECTS	: 5
Volume horaire de travail élève (enseignements + travail personnel)	: 132h
Nombre d'heures d'enseignement	: 63h

### Finalité de l'UE :

L'UE a pour but d'initier les étudiants aux approches multi-omiques et à la modélisation en épidémiologie.

Les approches multi-omiques visent l'intégration de données quantitatives obtenues pour différentes molécules biologiques afin de comprendre leurs interactions et d'avoir une image plus complète du phénotype biologique. Après avoir revue les bases de la biologie moléculaire, les différentes méthodes d'association entre un génotype et un trait allélique sont étudiées. [... la suite ...]

L'épidémiologie a pour objet l'étude des causes des maladies infectieuses et non infectieuses, de faire progresser et d'évaluer les méthodes de prévention des maladies et d'aider à planifier et à évaluer l'efficacité des programmes de santé publique.

### Structuration de l'UE :

Analyse de données Omics	36 h	2,5 ECTS
Épidémiologie	15 h	1,5 ECTS
Modélisation compartimentale	12 h	1 ECTS

L'épidémiologie s'intéresse à la fréquence et à la distribution des états de santé entre populations et individus. Elle cherche à identifier les déterminants des maladies afin de mettre en œuvre des mesures de prévention, de contrôle ou de soins efficaces.

Les approches « omiques » sont un ensemble d'outils analytiques permettant d'explorer les fonctions de notre organisme, dont l'une des plus connues est la génomique. Ces analyses permettent d'identifier des biomarqueurs prédictifs ou pharmacodynamiques qui seront d'une importance capitale dans le cadre de la compréhension des maladies ainsi que dans le cadre de la recherche clinique.

### Compétences ou acquis d'apprentissage à l'issue de l'UE :

- Décrire un événement épidémiologique et en investiguer les causes potentielles dans une logique de prévention ou de contrôle.
- Adapter les connaissances spécifiques au virus responsable pour mesurer l'efficacité des stratégies de lutte contre l'épidémie.
- Expliquer l'intérêt des données omiques dans la recherche clinique et épidémiologique.
- Conduire l'analyse des données omiques via des méthodes d'apprentissage statistique.

### Les pré-requis de l'UE :

Probabilité, statistique inférentielle, Statistique exploratoire multivariée, SAS, R (1A)  
Régression linéaire, GLM, Apprentissage supervisé, Modèle de durée, Chaîne de Markov, Calcul Bayésien, Méthodes de rééchantillonnage (2A)

UE 5 – Épidémiologie &amp; Données Omiques

## ÉPIDÉMIOLOGIE

### *Epidemiology*

<i>Enseignants</i>	: Olivier GRIMAUD (EHESP) et Cindy PADILLA (EHESP)
<i>Nombre d'ECTS</i>	: 1,5
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 30h
<i>Répartition des enseignements</i>	: 15h de cours
<i>Langue d'enseignement</i>	: Français
<i>Logiciels</i>	: /
<i>Documents pédagogiques</i>	: sur Moodle
<i>Pré-requis</i>	: Probabilité, Statistique inférentielle, SAS, R (1A) Régression, GLM (2A)

L'épidémiologie s'intéresse à la fréquence et à la distribution des états de santé entre populations et individus. Elle cherche à identifier les déterminants des maladies afin de mettre en œuvre des mesures de prévention, de contrôle ou de soins efficaces.

### Principales notions abordées :

L'épidémiologie en tant que science de santé publique

Indicateurs de fréquence

Schémas d'étude en épidémiologie et indicateurs d'association : lien entre tabagisme et cancer du poumon

Biais et facteur de confusion

Surveillance et investigation épidémiologique

Association et causalité

### Acquis d'apprentissage (objectifs) :

- Construire les indicateurs de fréquence que sont la prévalence et l'incidence.
- Choisir un schéma d'étude adapté à l'investigation d'un phénomène épidémiologique.
- Construire les indicateurs d'association utilisés pour identifier les facteurs de risque/déterminants
- Interpréter les résultats d'études épidémiologiques en tenant compte des sources d'erreurs.

### Modalités d'évaluation :

Examen écrit

### Références bibliographiques :

M.BERTIN, A FONTANET, Glossaire d'épidémiologie, 50 notions pour comprendre la pandémie de Covid-19. Presse de l'EHESP, 2021

R. BHOPAL, Concepts of Epidemiology. Integrating the ideas, theories, principles and method of epidemiology. OUP, 2016

K.J. ROTHMAN, S. GREENLAND, Modern epidemiology, Little, Brown and Company, Boston, 1998

UE 5 – Épidémiologie &amp; Données Omiques

## MODELISATION COMPARTIMENTALE

### *Compartmental Modeling*

<i>Enseignant</i>	: Audrey LAVENU (INSERM-CNRS)
<i>Nombre d'ECTS</i>	: 1
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 30h
<i>Répartition des enseignements</i>	: ● Cours : 6h ● Atelier : 6h
<i>Langue d'enseignement</i>	: Français
<i>Logiciels</i>	: R
<i>Documents pédagogiques</i>	: sur Moodle
<i>Pré-requis</i>	: Probabilité, statistique inférentielle, SAS, R (1A) Régression linéaire, GLM, Modèle de durée, Calcul Bayésien, Méthodes de rééchantillonnage (2A) Statistique des Processus, Epidémiologie (3A)

En épidémiologie, les modèles Susceptible-Infectieux-Removed (SIR) fournissent un cadre de modélisation adapté pour décrire la dynamique de transmission des maladies infectieuses. Ces modèles constituent un outil important pour découvrir les mécanismes générant la dynamique des maladies observées, évaluer les stratégies de contrôle potentielles et prédire les futures flambées. Le cours fournit une introduction de base dans le développement de modèles de type SIR. En particulier, les problèmes couramment rencontrés lors de la structuration et de l'analyse de ces modèles seront mis en évidence.

### Principales notions abordées :

Contexte de la modélisation compartimentale. Définition d'une épidémie.  
Modèle SIR déterministe  
Calcul du taux de reproduction de base ( $R_0$ )  
Simulation d'épidémies  
Exemples de données de surveillance épidémiologique  
Exemples de modèles dérivés du modèle SIR.  
Modèle SIR stochastique

### Acquis d'apprentissage (objectifs) :

- Décrire les principaux modèles compartimentaux en épidémiologie
- Simuler des épidémies par des modèles Susceptible-Infectieux-Removed (SIR)
- Interpréter les différents paramètres pour le modèle SIR et ses dérivés afin de guider la prise de décision en matière de santé publique.
- Traduire des hypothèses de propagation d'épidémies en système d'équations différentielles
- Simuler des modèles stochastiques et représenter un comportement moyen
- Simuler des stratégies de lutte contre les épidémies et mesurer leur efficacité

### Modalités d'évaluation :

Dossier écrit (10h de travail)

### Références bibliographiques :

- ANDERSON, R. A. and MAY R. M. 1982. *Directly transmitted infectious diseases: Control by vaccination*, Science 215, 1053-1060.
- ANDERSON R. A. and MAY R. M., 1992: *Infectious Diseases of Humans: Dynamics and Control* (2nd ed.). Oxford University Press, Oxford.
- BARTLETT M. S. 1960. *Stochastic Population Models in Ecology and Epidemiology*, Methuen, London.
- Y LI, Michael. *An introduction to mathematical modeling of infectious diseases*. Springer, 2018.

UE 5 – Épidémiologie &amp; Données Omiques

## ANALYSE DES DONNEES OMIQUES

### *Omic Data Analysis*

<i>Enseignant</i>	: Guillaume DESACHY (AstraZeneca), Pierre NEUVIAL (IMT), Isabel BRITO (Institut Curie)
<i>Nombre d'ECTS</i>	: 2,5
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 72h
<i>Répartition des enseignements</i>	: - Cours : 18h - Atelier : 18h
<i>Langue d'enseignement</i>	: Français
<i>Logiciels</i>	: R (knitr, tidyverse, ggplot2, caret, tidymodels, factoextra, cIValid, pROC, Bioconductor)
<i>Documents pédagogiques</i>	: Polycopié
<i>Pré-requis</i>	: Probabilité, statistique inférentielle, Statistique exploratoire multivariée, R (1A) Régression linéaire, GLM, Apprentissage supervisé, Chaîne de Markov, Méthodes de rééchantillonnage (2A) Statistique des Processus, Apprentissage Automatique (3A)

Les approches « omiques » sont un ensemble d'outils analytiques permettant d'explorer les fonctions de notre organisme. L'une des plus connues est la génomique, qui correspond à l'étude du génome, c'est à dire à tout ce qui touche à l'ADN, ses modifications, ses mutations, son expression. Ces analyses permettent d'identifier des biomarqueurs dits pronostiques, prédictifs ou pharmacodynamiques, qui seront d'une importance capitale dans le cadre de la compréhension des maladies ainsi que dans le cadre de la recherche clinique. Le cours consiste en une introduction à la Génomique et Post-Génomique (Transcriptomique, Protéomique). Les trois grandes familles d'analyse pour les données -omiques, à savoir *Class Discovery* (1), *Class Comparison* (2) et *Class Prediction* (3) seront présentées couvrant un large spectre d'applications en données -omiques.

### Principales notions abordées :

Présentation des données en épidémiologie génétique, et des données omiques en recherche clinique.

(1) *Class Comparison* : GWAS, Analyse Différentielle pour les données d'expression - méthodes post hoc,

(2) *Class Discovery* : Analyse de données de nombre de copies d'ADN - segmentation, Indices d'évaluation, Modèle de mélange, Réseau de gènes

(3) *Class Prediction*.

### Plan du cours :

Guillaume DESACHY (Cours 6h / Atelier 6h)

- Introduction à la génomique, Présentation des données en épidémiologie génétique, et des données omiques en recherche clinique.
- (1) *Class Comparison* : GWAS
  - Test du Chi-deux, Test exact de Fisher, Test d'égalité des moyennes, Test de Wilcoxon, Test multiples, Manhattan plot
  - Régression logistique, Régression linéaire multiple, Analyse en Composantes Principales
  - Equilibre de Hardy-Weinberg
- (2) *Class Discovery* : Classification ascendante hiérarchique, k-means, Indices de Rand, indice de Jaccard
- (3) *Class Prediction* : régression logistique et régression linéaire simple, Régression pénalisée (LASSO/elasticNet), Forêts aléatoires (RF)

Pierre NEUVIAL (Cours 6h / Atelier 6h)

- (1) *Class Comparison* : Analyse Différentielle pour les données d'expression - méthodes post hoc (3h)
  - Tests multiples, Inférence post hoc par procédure de test fermée
- 
- (2) *Class Discovery* : Analyse de données de nombre de copies d'ADN – segmentation (3h)
  - Modèles de mélange, moyenne mobile
  - Programmation dynamique, points de rupture

Isabel BRITO (Cours 6h / Atelier 6h)

- (1) *Class Comparison* : Analyse Différentielle de données d'expression - RNA-seq (3h)
  - Test exact de Fisher, Modèle de Poisson, Modèle Binomial Négatif, Test multiples
- (2) *Class Discovery* : Réseaux de gènes (6h)
  - Réseaux de régulation des gènes, réseaux sans échelle, données d'expression de gènes utilisées pour découvrir la structure du réseau
  - Modélisation des relations entre variables, graphes d'indépendance conditionnelle
  - Inférence de réseau pour les données d'observation
  - Modèle graphique gaussien, réseaux d'association, corrélation partielle, Réseaux de corrélation partielle (GGM)
- (2) *Class Discovery* : Analyse de la co-expression de gènes, modèle de mélange (1h30)

**Acquis d'apprentissage (objectifs) :**

- Expliquer le fonctionnement d'un organisme à l'échelle du génome (l'ADN, l'ARN, la protéine, la cellule).
- Identifier les approches omiques en épidémiologie et en recherche clinique.
- Analyser les associations entre un génotype et un trait via un modèle linéaire généralisé ou via des mesures d'association allélique.
- Appliquer une procédure de tests multiples ou des inférences post-hoc sur des données génomiques.
- Analyser des données -omiques à l'aide de procédures d'apprentissage supervisée et non supervisée

**Modalités d'évaluation :**

Trois QCMs à la fin de chaque partie (la moyenne des trois notes est coefficient 1), ainsi qu'une présentation orale sous la forme d'un poster (coefficient 2).

**Références bibliographiques :**

FOULKES, A. *Applied statistical genetics with R*. Springer, 2009.  
 LAIRD, N. and LANGE, C. *The fundamentals of modern statistical genetics*. Springer, 2010.  
 ZHENG, G. *Analysis of genetic association studies*. Springer, 2012.  
 GONDRO, Cedric. *Primer to analysis of genomic data using R*. Springer, 2015.  
 DRAGHICI, Sorin. *Statistics and data analysis for microarrays using R and bioconductor*. CRC Press, 2016

UE Spécifiques filière Santé &amp; Biostatistique

## UE : PROJETS

<i>Correspondant de l'UE</i>	: Myriam VIMOND
<i>Nombre d'ECTS</i>	: 5
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: Entre 120h et 150h
<i>Nombre d'heures d'enseignement</i>	: Suivis réguliers avec les encadrants (58h)

### Finalité de l'UE :

Les projets sont l'occasion pour les étudiants de mettre en œuvre leurs connaissances acquises à l'ENSAI sur des cas d'études concrets. Ils visent à mettre en œuvre les outils et connaissance acquises en statistique, en informatique et en économie, dans une démarche de résolution de problèmes concrets type ingénieur.

Les projets se déclinent en deux versions: le projet méthodologique, en langue anglaise, vise à approfondir une thématique centrée autour d'un ou plusieurs articles scientifiques ; le projet de fin d'études (dans toutes les filières sauf biostatistique), plus appliqué, nécessairement sur des données issues d'une collecte, vise à proposer une solution pratique à une problématique générale proposée par une entreprise ou un laboratoire de recherche. À eux deux, ces projets couvrent toute l'étendue d'une démarche de développement: diagnostique d'un problème nouveau, lecture de la littérature scientifique sur le sujet, résolution d'un problème en respectant un compromis entre les règles de l'art d'une part, et les contraintes humaines, financières et techniques de l'autre. Ils permettent par ailleurs aux élèves de mesurer l'utilité de toutes les notions acquises au cours des trois années de formation.

Selon les filières, la réalisation d'un Data Challenge complète ces cas d'études concrets, à travers la réalisation d'un projet sur un temps court et des contraintes spécifiques.

### Structuration de l'UE :

Projet méthodologique: approfondissement d'une démarche de revue de littérature en langue anglaise ; constitue la partie théorique de recherche d'information dans une démarche de recherche et développement. Les élèves désireux de développer un projet de recherche académique peuvent privilégier un projet plus conséquent.

Projet de fin d'étude: approfondissement d'une démarche pratique, sachant composer avec des contraintes opposées, entre rigueur scientifique et nécessités pratiques ; constitue la partie implémentation dans une démarche de recherche et développement.

Data Challenge (optionnel, selon les filières) : rassembler sur une période très courte différentes équipes de profils variés afin de collaborer sur un projet.

### Compétences ou acquis d'apprentissage à l'issue de l'UE :

Ces projets concluent la formation d'ingénieur de l'Ensaï, et mobilisent un ensemble de compétences de l'ingénieur : capacité à trouver l'information pertinente, à faire une veille scientifique, à prendre en compte les enjeux de l'entreprise, à travailler dans un contexte international, tout en mobilisant des compétences techniques pour résoudre des problèmes complexes, et mener une démarche scientifique.

### Les pré-requis de l'UE :

Méthodes de travail des projets de 1<sup>ère</sup> et 2<sup>ème</sup> année.

UE Projets

## PROJET METHODOLOGIQUE

### *Methodological project*

<i>Enseignants</i>	: Divers intervenants
<i>Nombre d'ECTS</i>	: 3
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 75h
<i>Répartition des enseignements</i>	: 10h d'ateliers, et suivis réguliers
<i>Langue d'enseignement</i>	: Anglais
<i>Logiciels</i>	: Sans objet
<i>Documents pédagogiques</i>	: Slides/Manuscrit
<i>Pré-requis</i>	: Sans objet

Le projet méthodologique est un projet académique, en langue anglaise, visant à approfondir une thématique centrée autour d'un ou plusieurs articles scientifiques. Au cours de ce projet, les étudiants auront généralement à approfondir une thématique et conduire par eux-mêmes une revue de littérature si besoin.

### Principales notions abordées :

Travail de recherche en groupe suivi par un chercheur (env. 5 séances) et un professeur d'anglais (4 séances).

### Acquis d'apprentissage (objectifs) :

- Démontrer la compréhension d'une méthodologie statistique soit par l'étude théorique de celle-ci ou soit par l'évaluation critique de celles-ci.
- Rédiger un article scientifique dans le domaine des statistiques appliquées ou fondamentales.
- Comparer différentes méthodologies statistiques pour un problème donné.
- Présenter de manière rigoureuse et précise une méthodologie ou des résultats.
- Employer le vocabulaire scientifique adéquat pour présenter un problème.
- Ecrire en anglais scientifique.
- Présenter oralement les résultats en anglais.

### Modalités d'évaluation :

Le choix des sujets a lieu dès la semaine de la rentrée. Le travail est effectué en binôme.

A l'issue du projet, chaque groupe d'élèves doivent produire un article scientifique en anglais, et une défense orale de leurs travaux en anglais également.

Le rapport doit être rendu pour fin janvier, la soutenance étant prévue dans les 10 jours qui suivent.

### Références bibliographiques :

HOTALING, Scott. Simple rules for concise scientific writing. *Limnology and Oceanography Letters*, 2020, vol. 5, no 6, p. 379-383.

LEE, Michael SY. Helpful Hints for Writing Your First Scientific Paper. 2020.

GEMAYEL, Rita. How to write a scientific paper. *The FEBS journal*, 2016, vol. 283, no 21, p. 3882-3885.

UE Projets

## PROJET DE FIN D'ETUDES

### *Final Year Project*

<i>Enseignants</i>	: Divers intervenants
<i>Nombre d'ECTS</i>	: 1
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 30h
<i>Répartition des enseignements</i>	: 18h d'ateliers, et suivis réguliers
<i>Langue d'enseignement</i>	: Français
<i>Logiciels</i>	: Sans objet
<i>Documents pédagogiques</i>	: Slides/Manuscrit
<i>Pré-requis</i>	: Sans objet

Le projet de fin d'études consiste en la production d'une étude statistique de niveau professionnel dans le monde de l'entreprise ou de la recherche, parmi un catalogue de sujet mis à disposition des élèves. Dans le cadre de la filière Biostatistique, ce projet aura la forme d'un datathon sur trois/quatre jours à la fin du mois de mars avec une thématique autour des données -omiques.

### Principales notions abordées :

Travail autonome en groupe avec des données -omiques fournie par un professionnel de l'entreprise ou de la recherche. Ce travail effectué sur une semaine fin mars.

### Acquis d'apprentissage (objectifs) :

- Définir une stratégie d'étude en réponse à une demande client
- Démontrer la compréhension d'une procédure statistique dans sa bonne application sur des données.
- Travailler en équipe
- Produire une étude en Sciences des Données qui soit reproductible

### Modalités d'évaluation :

Le projet est évalué à travers d'une soutenance orale au terme des trois/quatre journées.

Les étudiants devront produire un mémo de deux pages résumant leur étude, ainsi que les codes qui auront produit les résultats.

UE Projets

## DATA CHALLENGE

### *Data Challenge*

<i>Enseignants</i>	: Divers intervenants industriels (correspondante : Salima El Kolei)
<i>Nombre d'ECTS</i>	: Pas d'attribution d'ECTS
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 2 journées
<i>Répartition des enseignements</i>	: 12h d'ateliers
<i>Langue d'enseignement</i>	: Français
<i>Logiciels</i>	: Sans objet
<i>Documents pédagogiques</i>	: Slides/Manuscrit
<i>Pré-requis</i>	: Méthodes de travail des projets, Compétences statistiques et informatiques de 3ème année

### Modalités d'évaluation :

Les élèves participent au data challenge proposé à l'Ensaï ouvert également aux élèves de deuxième année. Il n'y a pas d'évaluation.

### Acquis d'apprentissage (objectifs) :

Le data challenge permet de rassembler sur une période très courte différentes équipes de profils variés afin de collaborer sur un projet. Cette expérience se rapproche des conditions réelles dans laquelle évoluent les datascientists au sein des entreprises. Il permet, à partir des mécanismes du jeu, de dynamiser et d'articuler la pédagogie autour d'un besoin concret d'entreprise et d'un événement qui s'achève par une évaluation objective. De nombreux challenges sont proposés autour de la Data ou présentant des problématiques Data importantes.

L'objectif de ce cours est de valoriser les compétences transversales acquises dans ce contexte opérationnel. Les compétences qui sont renforcées grâce à celui-ci sont multiples:

- Comprendre les problèmes à résoudre.
- Travailler en mode projet avec des contraintes.
- S'intégrer et s'adapter dans un contexte pluridisciplinaire. Selon les challenges, les compétences seront mobilisées à géométrie variable.
- S'adapter à la réalité de la Data d'entreprise (données non structurées, manquantes, volumétrie...)
- Communication orale des résultats (pitch...)

### Principales notions abordées :

Travail en groupe sur un temps court.

UE Projets

## ANGLAIS

### English

<i>Enseignants</i>	: Divers intervenants (correspondant : Todd Donahue)
<i>Nombre d'ECTS</i>	: 1
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 45h
<i>Répartition des enseignements</i>	: 15h de cours, 15h d'aide au projet de fin d'études
<i>Langue d'enseignement</i>	: Anglais
<i>Logiciels</i>	: Sans objet
<i>Documents pédagogiques</i>	: Sous Moodle
<i>Pré-requis</i>	: Aucun

### Modalités d'évaluation :

L'examen final prend la forme d'une simulation d'entretien d'embauche. Cet examen oral durera environ 25 minutes, sera noté, et permettra d'évaluer le niveau d'expression orale sur l'échelle CECRL (Cadre européen commun de référence pour les langues). Le CV et la lettre faite pour cet exercice seront évalués et feront partie de la note finale. L'anglais est également évalué à travers le rapport écrit et la soutenance orale du projet de fin d'études. Le niveau acquis apparaîtra sur le Supplément au diplôme. L'objectif de la CTI (Commission des Titres d'Ingénieur) pour tous les élèves ingénieurs est d'atteindre le niveau B2.

### Acquis d'apprentissage (objectifs) :

- maîtriser une ou plusieurs langues étrangères
- savoir candidater et réussir un recrutement en langue anglaise
- contextualiser et prendre en compte les enjeux et les besoins de la société
- se connaître, s'auto-évaluer, gérer ses compétences, opérer ses choix professionnels
- s'intégrer et évoluer dans un groupe pour mener à bien un projet dans un contexte international et/ou pluriculturel
- savoir identifier les informations pertinentes, à les évaluer et à les exploiter

### Principales notions abordées :

Pour les élèves qui n'ont pas eu un score d'au moins 785 au TOEIC : pendant les 5 premières séances, la plupart des cours seront basés sur la préparation à cet examen. Les ressources informatiques de l'Ecole doivent aussi être mises à profit (pages Moodle, TOEIC Mastery), ainsi que les méthodes disponibles à la bibliothèque. Pour les autres élèves, les cours seront organisés par groupe de niveau et conçus afin de les préparer à affronter le monde professionnel sur le plan international. Les thèmes suivants seront traités : « Leading meetings », « Interviews », « Presentations », « Taking decisions », et « Negotiating deals », et « Cultural and Political Current Events ». Ensuite, les 5 dernières séances seront consacrées au travail de rédaction/correction des rapports faits en anglais dans chaque filière ainsi qu'à la préparation des soutenances orales. Chaque responsable de filière indiquera aux élèves, en début d'année, le projet concerné et les modalités de notation. Les élèves recevront des consignes détaillées avant de démarrer ces cinq séances, afin d'arriver à la première séance avec une première version ou extrait de leur rapport en anglais prêt pour correction et relecture. Pour tout complément d'information, chaque élève peut consulter le Programme des enseignements : Langues étrangères, distribué au début de l'année académique.

Références bibliographiques : Définies par chaque intervenant.

UE Stages

## UE : PROJET PROFESSIONNEL ET STAGES

<i>Correspondant de l'UE</i>	: Patrick Gandubert
<i>Nombre d'ECTS</i>	: 30
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: Travail en entreprise
<i>Nombre d'heures d'enseignement</i>	: 30h (séminaires)

### Finalité de l'UE :

Cette UE correspond à des temps pédagogiques en lien direct avec les entreprises. Les séminaires professionnels ont pour objectif de présenter aux étudiants diverses problématiques auxquelles ils seront confrontés dans leur environnement professionnel. Il permet d'apporter des compléments par rapport à certains cours, et fait le lien entre les enseignements et les applications pratiques qui en découlent. Le projet professionnel permet de préparer les étudiants à leur entrée dans la vie professionnelle et aux stages, il est réalisé sur la 2ème et 3ème année de formation. Des simulations d'entretien de recrutement sont organisées en 3e année. Elles sont assurées par des recruteurs d'entreprises et d'organisations partenaires de l'Ensaï. Les stages (application en 2ème année, fin d'études en 3ème année) permettent aux élèves de mettre en pratique les enseignements de mathématiques appliquées, d'informatique et d'économie dans un cadre professionnel. Le stage de fin d'études, d'une durée de 20 semaines minimum, vise à appliquer les enseignements de 3ème année et à acquérir de l'expérience pour assurer la transition vers l'emploi. Il constitue une étape essentielle de mise en situation professionnelle pour le futur ingénieur qui dispose à ce stade de l'ensemble des bagages techniques de la formation.

### Structuration de l'UE :

Le stage de fin d'études constitue la majeure partie de l'évaluation de cette UE (25 ECTS). L'Ensaï exige une forte adéquation entre le contenu du stage et la filière de spécialisation de 3e année. Il fait l'objet d'une procédure de validation par le responsable de filière et par le département des relations avec les entreprises. L'évaluation tient compte de la capacité d'intégration de l'étudiant dans l'entreprise, ses capacités d'initiative et de satisfaction au regard des objectifs du stage, et de la qualité du rapport et de la soutenance réalisée devant un jury composé d'un président, d'un vice-président, tous les deux issus du monde de l'entreprise et d'un permanent de l'école. Le stage d'application de 2ème année est pris en compte dans cette UE (5 ECTS). Les séminaires professionnels ne sont pas évalués.

### Compétences ou acquis d'apprentissage à l'issue de l'UE :

Le stage de fin d'études (et l'UE) comprend un objectif technique - il s'agit de répondre à la commande, à la problématique inscrite dans le thème du stage à l'aide des connaissances acquises - et un objectif professionnel - il s'agit de parfaire la connaissance du monde du travail, de développer des capacités relationnelles et d'adopter une démarche d'insertion dans le monde professionnel.

### Les pré-requis de l'UE :

Aucun

UE Projet Professionnel

## DROIT DU TRAVAIL

### *Work Law*

<i>Enseignant</i>	: Charlotte GRUNDMAN, Avocat au Barreau de Paris
<i>Nombre d'ECTS</i>	: 0
<i>Volume horaire de travail élève (enseignements + travail personnel)</i>	: 15h
<i>Répartition des enseignements</i>	: Cours : 3h – Atelier : 6h
<i>Langue d'enseignement</i>	: Français
<i>Logiciels</i>	: Sans objet
<i>Documents pédagogiques</i>	: Distribués pendant le cours
<i>Pré-requis</i>	: Aucun

### Modalités d'évaluation :

Exposé d'un cas pratique réalisé lors des TD.

### Acquis d'apprentissage (objectifs) :

La matière étant extrêmement vaste et complexe, il est ici proposé aux étudiants une approche didactique et vivante du sujet, l'objectif de l'enseignement étant de permettre aux étudiants qui travailleront dans un futur proche en entreprise d'avoir compris certaines notions pratiques essentielles en droit du travail.

### Principales notions abordées :

Hormis le cours d'amphi, il sera systématiquement proposé aux étudiants, après l'étude d'une notion, un exercice visant à mettre en pratique la notion abordée. Le cours commun (3 heures) traite des notions suivantes : Comprendre d'où l'on vient pour savoir où on va (introduction historique au droit du travail, les sources du droit du travail, ordre public absolu et ordre public social), les instances de contrôle du droit du travail, formation et exécution du contrat de travail, la rupture du contrat à durée indéterminée. Pour les TD, la première heure de cours sera consacrée à l'étude d'un chapitre (la modification du contrat de travail, le recrutement, les droits fondamentaux du salarié). Cet exposé sera suivi d'une mise en situation pratique, où les étudiants devront par groupe répondre à un cas pratique. Un rapporteur sera désigné par groupe, et la notation se fera à cette occasion.

## Offre de Formation Par la Recherche

L'Ensaï a signé des conventions de double diplôme avec trois masters de l'Université de Rennes, correspondant aux différents domaines de la recherche en biostatistique : la statistique mathématiques, la recherche clinique, l'épidémiologie, la pharmacologie et la génomique.

### Master de mathématiques fondamentales, Université Rennes 1

L'objectif principal est d'acquérir et de consolider un bagage mathématique en probabilités et statistiques. Selon la convention, les étudiants doivent valider les cours suivants : *Statistique des processus*, *Estimation paramétrique*, *estimation nonparamétrique*, le *Séminaire de master*.

UE	Dispense	Crédits
Statistique des processus	Statistique des Processus	6
Estimation paramétrique		6
Estimation nonparamétrique		6
Séminaire de Master	Projet Méthodologique	6
Traitement automatique du langage	<i>Cours Ensaï</i>	6
UE « Autre cours ENSAI » (à choisir)	<i>Cours Ensaï</i>	6
Anglais	<i>Cours Ensaï</i>	3
Stage	<i>Stage</i>	15

### Master Bio-informatique en santé, Université de Rennes 1

Ce parcours vise à se spécialiser sur les enjeux actuels de la génétique et de la génomique en santé : aide au diagnostic, génomique et cancérologie, génétique de l'hérédité simple et complexe, ingénierie génétique, ainsi que sur la conception de thérapeutiques nouvelles et personnalisées.

	Dispense	Crédits
UE « Bioinformatique en génomique » M1		6
UE « Biologie des systèmes : réseaux biologiques » M2		6
UE « Genemapping and NGS analysis » M2		6
UE « <i>Essais cliniques</i> » ENSAI	<i>Cours Ensaï</i>	6
UE « Introduction à l'analyse des données omiques & Analyse des données omiques » ENSAI	<i>Cours Ensaï</i>	6
UE Stage	<i>Stage ingénieur</i>	30

Dispense : « Droit des entreprises » et « Mesures de la qualité de vie ». Les crédits ECTS associés sont attribués sous condition que le Master Bioinformatique soit validé.

### Master de Santé publique, parcours Modélisation en Pharmacologie Clinique et Epidémiologie (MPCE), Université Rennes 1

Ce parcours vise à former des chercheurs ou des professionnels capables de concevoir et d'analyser de façon approfondie des données de tous types de protocoles de recherche clinique (pharmacologie clinique, essais thérapeutiques), épidémiologique et pharmaco-épidémiologique, et de développer une recherche méthodologique adaptée à ces différents domaines. Les étudiants doivent choisir 10 unités d'enseignements dans la liste suivante selon l'option choisie.

L'option « *Biostatistique* » (BS) forme les étudiants au développement et à la validation de modèles et méthodes statistiques appropriées pour l'analyse et l'interprétation de données complexes en épidémiologie ou en recherche clinique.

L'option « *Pharmacologie clinique et évaluation thérapeutique* » (PCET) forme les étudiants à la conception (méthodologie) et à l'analyse approfondie des données de tous types de protocoles de recherche clinique (pharmacologie clinique, essai thérapeutique). Les étudiants apprendront à synthétiser les résultats de plusieurs essais cliniques grâce à la démarche de la méta-analytique. Les étudiants seront également initiés à la modélisation pharmacocinétique et pharmacocinétique-pharmacodynamique par une approche individuelle et de population.

L'option « *Epidémiologie* » (EPI) forme les étudiants aux concepts et méthodes statistiques spécifiques aux études épidémiologiques à visée étiologique et clinique, aux études d'épidémiologie génétique, à l'évaluation de méthodes diagnostiques et à la recherche de facteurs pronostiques des maladies.

L'option « *Pharmaco-épidémiologie* » forme les étudiants à la méthodologie permettant d'évaluer les risques et les bénéfices des médicaments tels qu'ils sont utilisés par les patients, à l'analyse de l'impact de ces traitements en termes de santé publique, dans les conditions réelles d'usage des médicaments. Les étudiants seront formés à la gestion de grandes bases de données, particulièrement la base du système national d'information inter-régimes de l'Assurance maladie (SNIIRAM).

Cette année seul le parcours EPI serait ouvert selon le projet de convention suivant :

- **UE1 : Planification d'une étude de recherche biomédicale**  
*dispense de l'examen essais clinique (Ensaï)*  
*Notons que les étudiants devront suivre la partie projet du cours d'essais cliniques de l'Ensaï.*
- UE2 = Modèles Mixtes à l'Ensaï (24h, examen écrit)
- UE3 = Analyse de Données de Survie à l'Ensaï Survie (24h, examen écrit)
- **UE5 : Epidémiologie Clinique**  
*dispense du cours d'Epidémiologie (Ensaï)*
- UE10 = Statistique Bayésienne (18h,devoir maison), traitement de la donnée manquante (15h, TP noté) à l'Ensaï
- **UE7 : Recherche Etiologique**
- UE9 = Analyse de Données Omiques à l'Ensaï (36h)
- UE10 = Statistique Bayésienne (18h,devoir maison), traitement de la donnée manquante (15h, TP noté) à l'Ensaï
- **UE12 : Validation et analyse des échelles de mesures en santé**  
*dispense echelle de mesure en santé(Ensaï)*
- **UE13 : Méthodes Statistiques Avancées en Evaluation thérapeutiques**  
*dispense du cours de Méta-Analyse (Ensaï)*

Convention pour l'année 2020/2021 :

UE	BS	EPI	PEPI	PCET	Crédits
UE 1 « Planification d'une étude biomédicale »	X	X	X	X	3
UE2 « Statistique des processus & Statistique Bayésienne »	<i>Cours Ensai</i>				3
UE3 « Plans d'expériences & Modèles Mixtes »	<i>Cours Ensai</i>				3
UE4 « Méthodes statistiques avancées en pharmacologie clinique »	X		X	X	3
UE5 « Méthodes statistiques avancées en épidémiologie » (Dispense EPI et PEPI – Epidémiologie)	X	X	X		3
UE6 « Principe d'estimation en biostatistique » ou « Modèle de durée & Modélisation compartimentale & Traitement des données manquantes »	<i>Cours Ensai</i>	opt	opt	<i>Cours Ensai</i>	3
UE7 « Recherche étiologique »	opt	X	opt		3
UE8 « Epidémiologie clinique »	opt	X	X		3
UE9 « Analyses Exploratoires Multidimensionnelles »	opt	opt	opt		3
UE10 « Validation et analyse des échelles en mesures de santé » (Dispense BS – Mesure de qualité de vie) ou « Modèle de durée & Modélisation compartimentale & Traitement des données manquantes »	X	<i>Cours Ensai</i>		opt	3
UE 11 « Pharmaco-épidémiologie »	opt	opt	X	opt	3
UE12 « Méthodes en phase I et phase II »				X	
UE13 « Méthodes en pharmacocinétique et modélisation PK-KD » (Dispense PCET – Pharmacométrie)				X	
UE14 « Méthodes de statistiques avancées en éval. Thérapeutique » ou « Modèle de durée & Modélisation compartimentale & Traitement des données manquantes »	X	X	<i>Cours Ensai</i>	X	3
UE15 « Concepts et méthodes en chronopharmacologie »				opt	3
UE16 « Application de la modélisation PK-PD »				opt	3
UE option « Introduction à l'analyse des données omiques & Analyse des données omiques »	<i>Cours Ensai</i>				3
UE Stage	<i>Stage Ingénieur</i>				30