Programme des enseignements de 3e année

Filière SSV

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

FILIÈRE STATISTIQUE POUR LES SCIENCES DE LA VIE

ANNÉE SCOLAIRE 2020/2021

BIOSTATISTICS SPECIALIZATION

2020/2021 ACADEMIC YEAR

Présentation de la filière

La formation d'ingénieur de l'Ensai inclut 6 filières de spécialisation. Toutes ces filières forment aux métiers de la Data Science, avec une maîtrise des outils permettant l'extraction, l'analyse et la fouille de données et une capacité à choisir les modalités de traitements des données massives (Big Data) et des techniques d'apprentissage automatique (machine learning). Selon les spécialisations, ces compétences sont spécifiques à un domaine ou transversales. L'ensemble des filières continue à former aux compétences transversales (soft skills) et à la valorisation des travaux menés dans un contexte professionnel et international. Lors des cours et du projet méthodologique en anglais, les élèves travaillent toutes les compétences linguistiques et communicationnelles et approfondissent leurs connaissances liées au monde de l'entreprise et de la recherche. La séquence de Tronc Commun mêlant enseignements scientifiques, projets et anglais conclut la formation à l'autonomie et la capacité à mettre en œuvre des analyses de données en situation complexe. Un stage de fin d'études est à réaliser à l'issue de la scolarité, qui permet de mettre en œuvre dans un cadre professionnel une démarche scientifique autour d'une problématique en lien avec les enseignements de la filière.

La spécialité *Statistique pour les Sciences de la Vie* 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 de la science expérimentale et de la génétique, des essais cliniques et des études observationnelles. Les domaines d'application sont la médecine, l'agriculture, la foresterie, les sciences de l'environnement et au-delà.

L'objectif de cette spécialité est de former des biostatisticiens. Ce programme intègre une unité d'enseignement (UE) sur l'**apprentissage automatique** commune à toutes les spécialisations, et deux UEs spécifiques ayant pour objet d'approfondir et compléter les outils de **modélisation statistique** étudiés durant les deux premières années à l'Ensai. Ces deux UEs spécifiques comprennent l'étude des modèles mixtes, des modèles de survie, des plans d'experience, des processus stochastiques et l'inférence bayésienne pour des modèles hiérarchiques. En outre trois domaines importants de la biostatistique sont couverts comme l'**épidémiologie**, les **essais cliniques** et **la génomique**. Transversalement à ces unités d'enseignement, les applications en informatique (R, Python, SAS, WINBUGS, JAGS, etc) sont omniprésentes.

Ainsi les compétences visées à l'issu de cette année sont :

- la maîtrise des 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 « Omiques » (études de données génomiques, transcriptomiques, métabolomiques, protéomiques, épigénétiques et métagénomiques)
- la capacité à utiliser le raisonnement et la théorie probabiliste et statistique pour analyser des problèmes non standard survenant en médecine et en santé publique
- la 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).
- La capacité à choisir les modalités de traitements des données massives (Big Data) et des techniques de *machine learning* en santé.

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.

Option Formation Par la Recherche

L'Ensai offre la possibilité, aux élèves de 3ème 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

A l'issue de ce parcours, ils pourront poursuivre cette formation par une thèse académique ou de type Cifre (Convention Industrielle de Formation par la Recherche). Les thèses académiques sont en général encadrées dans des laboratoires de recherche tels que ceux du CNRS ou de l'Inserm. En ce qui concerne les entreprises signataires de thèses Cifre ou organismes de recherche, on peut citer par exemple l'I.R.I.S. –Laboratoires Servier associé à l'INSERM ou encore l'IRSN associé à l'Université de Paris 11.

Les entreprises partenaires

La filière bénéficie de partenariats avec des acteurs économiques de premier plan. Ces partenariats permettent de développer des échanges privilégiés notamment via des cours, des séminaires professionnels et des stages.

Amaris

INSEP
Terre de Champions

Inserm
Institut national
de la santé et de la recherche médicale

QUINTILES

Keyrus
BIOPHARMA

SANOFI

SERVIER

SOLADIS

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

Programme

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

	Volume	Crédits
UE0 Cours d'ouverture	69 h	2
Droit du Travail	9 h	0.5
Anglais	30 h	1.5
Sport	30 h	0
UE1 Apprentissage Automatique	78 h	6
Apprentissage automatique	27 h	2
Régression pénalisée et sélection de modèles	15 h	1
Apprentissage à grande échelle	15 h	1.5
Traitement automatique du langage	21 h	1.5
UE2 Méthodologie statistique 1	60 h	4
Plans d'expériences	18 h	1
Modèles mixtes	21 h	1.5
Modèles de durée	21 h	1.5
UE3 Méthodologie statistique 2	75 h	5
Mesures de qualité de vie	15 h	1
Statistique des processus	15 h	1
Traitement des données manquantes	12 h	1
Statistique bayésienne	15 h	1
Méta-analyse	18 h	1
UE4 Essais cliniques	60 h	4
Essais cliniques	18 h	1
Pharmacométrie	18 h	1
Projet d'essais cliniques	24 h	2
UE5 Épidémiologie / Génomique	57 h	4
Épidémiologie quantitative	15 h	1
Modélisation compartimentale	12 h	1
Introduction à l'analyse des données Omiques	12 h	1
Analyse de données Omiques	18 h	1
UE6 Projets	84 h	5
Projet méthodologique	36 h	2,5
Projet de fin d'étude	36 h	2,5
Data Challenge	12 h	0
Séminaires professionnels	30 h	0

UE 7 Projet professionnel et stages		**30**
Stage 2A		5
Stage 3A		25
TOTAL	513	60

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

Descriptifs des enseignements communs

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

UE 0

UE : COURS D'OUVERTURE

Correspondant de l'UE : Ronan Le Saout
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)
Nombre d'heures d'enseignement : 2
: Entre 50h et 60h

: 39h

Finalité de l'UE :

À la fin de cette UE, notamment grâce à l'enseignement de l'anglais, les élèves seront capables de mettre en œuvre les compétences linguistiques et culturelles qui facilitent la suivie des cours scientifiques dispensés en anglais ou d'autres langues, le travail dans un environnement professionnel international et la compréhension des normes culturelles dans les pays étrangers. À travers le cours de droit du travail, les élèves acquerront les connaissances dans une discipline autre que la statistique, l'économie et l'informatique nécessaires pour mieux appréhender le contexte juridique de l'entreprise. Cette UE vise également le développement des compétences transversales (*soft skills)* qui aideront les élèves à réussir les projets académiques de leur formation, à intégrer le marché du travail et à devenir les citoyens éclairés.

Structuration de l'UE :

L'UE 0 de la 3ème année se compose de deux matières obligatoires, l'anglais et le droit du travail, ainsi que le sport de manière optionnelle.

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

En anglais les élèves travaillent toutes les compétences linguistiques pour atteindre le niveau B2 du CECR et progresser vers un niveau C1. Lors des cours d'anglais et de l'aide au projet en anglais, les élèves développement également les connaissances liées au monde de l'entreprise et de la recherche ainsi que les compétences transversales *(soft skills).* Le cours du droit du travail permet aux étudiants d'identifier et comprendre certaines notions pratiques essentielles en gestion des ressources humaines en entreprise.

Les pré-requis de l'UE :

Aucun

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

UE 0 Cours d'ouverture

ANGLAIS

English

Enseignant : Divers intervenants (correspondant : Todd Donahue)
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)
Répartition des enseignements
Langue d'enseignement : 1
: 40h

: 15h de cours, 15h d'aide au projet de fin d'études
: Anglais
Logiciels : Sans objet
Documents pédagogiques : Sous Moodle
Pré-requis : 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
- 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 « Negociating 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 Cours d'ouverture

DROIT DU TRAVAIL

Work Law

Enseignant : Charlotte GRUNDMAN, Avocat au Barreau de Paris
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)

Répartition des enseignements
Langue d'enseignement : 1
: 15h

: Cours : 3h • Atelier : 6h
: Français
Logiciels : Sans objet
Documents pédagogiques : Distribués pendant le cours
Pré-requis : 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.

UE Cours d'ouverture

SPORT

Sport

Enseignant : Divers intervenants (correspondant : Jullien Lepage)
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel) : 0
: 30

Modalités d'évaluation :

La participation à une activité sportive peut donner lieu à l'attribution d'un bonus ajouté sur la moyenne du semestre concerné. Le niveau de ce bonus est précisé dans une circulaire d'application en début d'année académique. Il varie selon l'assiduité aux séances, l'engagement et la participation aux compétitions tout au long de l'année. Pour être définitive, la liste des élèves bénéficiant de ces bonus doit être validée par le directeur des études.

Un bonus peut être exceptionnellement attribué en dehors des activités sportives réalisées dans le cadre Ensai. Pour y prétendre, les élèves concernés doivent remplir les 3 conditions suivantes:
- pratiquer régulièrement une activité sportive et participer aux compétitions liées ;
- posséder un niveau national (voir très bon niveau régional suivant le sport en question) ;
- déposer une demande argumentée auprès de la direction des études et du service sport en début d'année scolaire, afin de faire valider le programme d'entraînement, des compétitions et les modalités de diffusion des performances. Pour certains ayant des contraintes sportives, des aménagements horaires pourront d'ailleurs être ainsi envisagés si besoin.

Acquis d'apprentissage (objectifs) :

L'objectif est d'amener les élèves à maintenir un esprit sportif, sortir du strict cadre académique et développer leurs capacités physiques.

Principales notions abordées :

Neuf activités sportives sont proposées par l'école : Badminton, Basket, Football, Hand-ball, Tennis de table, Tennis débutant, Volley-ball, Cross-training, Course à pied/préparation physique/coaching sportif. Outre les entraînements, les élèves inscrits peuvent être amenés à participer à des compétitions.

UE 1

UE : APPRENTISSAGE AUTOMATIQUE (MACHINE LEARNING)

Correspondant de l'UE : Arthur Katossky
Nombre d'ECTS

Volume horaire de travail élève
(enseignements + travail personnel)
Nombre d'heures d'enseignement : 6 pour les filières ISTS, GDR et SSV, 7 pour GS et MQRM, 8 pour SID
: De 25h à 30h par ECTS

: 78h pour les filières ISTS, GDR et SSV, 102h pour GS et MQRM, 120h pour SID

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 automatique 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 4 matières : apprentissage automatique (Machine-learning), régression pénalisées et régularisation, apprentissage statistique à grande échelle, 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…). Selon les filières de spécialisation, des séminaires complémentaires (systèmes de recommandation…) sont introduits.

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

Cette UE permet de maitriser 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, s'intégrant dans une architecture informatique de données volumineuses (big data).

Les pré-requis de l'UE :

Modélisation statistique de 2ème année, méthodes d'optimisation et d'algorithmique, panorama du big data, aisance en R et Python.

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

UE Machine Learning

APPRENTISSAGE AUTOMATIQUE

Machine Learning

Enseignant : Hong-Phuong DANG (Ensai), Romaric GAUDEL (Ensai), Fabien NAVARRO (Ensai) et Brigitte GELEIN (Ensai)
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)

Répartition des enseignements

Langue d'enseignement : 2 (ISTS, GDR, SSV), 3 (GS, MQRM) ou 4 (SID)
: De 25h à 30h par ECTS

: Pour les 27h en filières ISTS, GDR et SSV, il y a 15h de cours et 12h d'ateliers. Pour les 24h complémentaires en GS et MQRM, il y a 6h de cours et 12h d'ateliers. Pour les 36h complémentaires en SID, il y a 12h de cours et 18h d'ateliers.
: Français
Logiciels : R et Python
Documents pédagogiques : supports de cours, bibliographie et fiches de TP
Pré-requis : R, Python, modélisation statistique, algèbre linéaire, optimisation de fonctions

Modalités d'évaluation :

Contrôle continu à la discrétion des intervenants, un QCM, compte-rendus de TP (1 en ISTS, GDR et SSV, 4 en GS et MQRM, 4 en SID)

Acquis d'apprentissage (objectifs) :

- Identifier comment résoudre une tâche par apprentissage automatique
- Choisir un modèle a priori adapté à une tâche
- Utiliser un modèle de l'état de l'art (SVM, réseau de neurones, forêt, …)
- Comparer empiriquement différents modèles pour une tâche donnée

Principales notions abordées :

Un rappel des principes de l'apprentissage statistique et automatique sera effectué. L'ensemble des filières aborderont les réseaux de neurones (y compris deep learning), les méthodes d'agrégation (forêts aléatoires, bagging, boosting, stacking) et les séparateurs à vaste marge (Support Vector Machines). Les réseaux de neurones avancés seront abordés en GS, MQRM et SID, les systèmes de recommandation en MQRM et SID.

Références bibliographiques :

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

- Endrew Ng. Machine Learning Yearning. Disponible gratuitement au lien https://www.deeplearning.ai/machine-learning-yearning/.
- Rémi Gilleron. Apprentissage machine - Clé de l'intelligence artificielle - Une introduction pour non-spécialistes. Ellipses.
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. 2016

UE Machine Learning

REGRESSION PENALISEE ET SELECTION DE MODELES

Penalized problems and model selection

Enseignant : Cédric HERZET (INRIA) & Clément ELVIRA (INRIA)
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)

Répartition des enseignements
Langue d'enseignement : 1
: 30h

: Cours : 9h • Atelier : 6h
: Français
Logiciels : Python
Documents pédagogiques : Supports de cours, Supports de TP, Bibliographie
Pré-requis : Optimisation, Python, Algèbre

Modalités d'évaluation :

Un examen sur table de 2 heures avec questions de cours et résolution de problèmes, 1 compte-rendu de TP

Acquis d'apprentissage (objectifs) :

- identifier les cas d'apprentissage statistique et les problèmes inverses où la régularisation est utile : comprendre quels sont les motivations et les objectifs de la pénalisation
- connaître les modèles de régularisation les plus courants et savoir quelles caractéristiques de reconstruction ils favorisent choisir une régularisation parmi les méthodes les plus courantes et estimer un modèle régularisé par une méthode de descente de gradient
- connaître et comprendre les différents types de problèmes d'optimisation et les algorithmes qui permettent de les résoudre numériquement

Principales notions abordées :

- Ingrédients principaux des problèmes inverses et d'apprentissage statistique + exemples pratiques
- Motivations et objectifs de la régularisation
- Types de régularisation et fonctions de régularisation couramment rencontrées
- Caractérisation des problèmes pénalisés: existence de solution, unicité, conditions d'optimalité.
- Méthodes numériques de résolution de problèmes d'optimisation
- Conditions théoriques de reconstruction correcte

Références bibliographiques :

- Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. 2015. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press.
- C. Bishop. Pattern recognition and machine learning. Springer-Verlag New York, 2006.
- S. Foucart and H. Rauhut. A mathematical introduction to compressive sensing. Applied and Numerical Harmonic Analysis. Birkhaüser, 2013.
- D. P. Bertsekas. Nonlinear Programming. Athena Scientific, USA, 2003.

UE Machine Learning

APPRENTISSAGE A GRANDE ECHELLE

Large scale machine-learning

Enseignant : Romaric Gaudel (ENSAI)
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)

Répartition des enseignements
Langue d'enseignement : 1,5
: 25h

: Cours : 6h • Atelier : 9h
: Français
Logiciels : Python
Documents pédagogiques : support de cours, fiches de TP
Pré-requis : panorama du big data, machine-learning, optimisation

Modalités d'évaluation :

2 quizz en contrôle continu, 1 TP noté

Acquis d'apprentissage (objectifs) :

Le passage à des bases de données à grande échelle modifie certains usages en apprentissage statistiques. Nous en verrons quelques exemples, avec les justifications théoriques sous-jacentes.

- objectif : être en mesure de choisir des approches appropriées pour un problème donné. Nécessite de
- objectif : connaître le comportement en termes de taille de stockage et/ou de temps de calcul et/ou de qualité d'estimation des approches présentées.

Principales notions abordées :

- la descente de gradient stochastique
- données et modèles parcimonieux
- le calcul distribué

Références bibliographiques :

- Introduction to High Performance Computing for Scientists and Engineers, Georg Hager, Gerhard Wellein, CRC Press, 2010
- Introduction to High Performance Scientific Computing, Victor Eijkhout , Edmond Chow, Robert van de Geijn, 2014
- Bekkerman, Ron, Mikhail Bilenko, and John Langford. n.d. *Scaling up Machine Learning: Parallel and Distributed Approaches*.
- The MIT Press. n.d. “Large-Scale Kernel Machines.” Accessed September 1, 2020. https://mitpress.mit.edu/books/large-scale-kernel-machines.

UE Machine Learning

TRAITEMENT AUTOMATIQUE DU LANGAGE ET FOUILLE DU WEB

Webmining et NLP

Enseignant : Guillaume Gravier (Irisa)
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)

Répartition des enseignements
Langue d'enseignement : 1,5 sauf pour filière informatique (SID) : 2,5
: 38h sauf pour filière informatique (SID) : 56h

: Cours : 9h • Atelier : 12h (6h complémentaires en SID)
: Français
Logiciels : Python
Documents pédagogiques : Support de cours, Supports de TP
Pré-requis : Python, Machine Learning

Modalités d'évaluation :

Projet (de taille plus importante en SID)

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).

UE Projets

UE : PROJETS

Correspondant de l'UE : Arthur Katossky
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)
Nombre d'heures d'enseignement : 5
: Entre 120h et 150h

: Suivis réguliers avec les encadrants

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 de déclinent en deux versions: le projet académique, en langue anglaise, vise à approfondir une thématique centrée autour d'un ou plusieurs articles scientifiques ; le projet de fin d'études, 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 rigoureuse, à la pointe de la recherche scientifique, en langue anglaise ; constitue la partie théorique de recherche d'information dans une démarche de recherche et développement.
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'Ensai, 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 1ère et 2ème année.

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

UE Projets

PROJET MÉTHODOLOGIQUE

Methodological project

Enseignant : Divers intervenants
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)

Répartition des enseignements
Langue d'enseignement : 2,5
: Entre 60h et 75 h

: 9h d'ateliers, et suivis réguliers
: Anglais
Logiciels : /
Documents pédagogiques : /
Pré-requis : /

Modalités d'évaluation :

Le projet méthodologique consiste en la production d'un article de synthèse sur un sujet de recherche à choisir parmi un catalogue. L'évaluation tient compte de l'article rédigé et de la réalisation d'une soutenance.

Acquis d'apprentissage (objectifs) :

Les objectifs du projet méthodologique, et donc les compétences qui sont renforcées grâce à celui-ci, sont multiples:

- familiarisation avec la forme des productions académiques (articles notamment), en lecture comme en écriture
- capacité à faire une revue de littérature mélangeant ouvrages scientifiques et professionnels
- mise en œuvre d'une démarche scientifique rigoureuse
- prise de conscience des enjeux autour de la reproductibilité des résultats de recherche
- communication sur des sujets techniques

À cela s'ajoute les objectifs spécifiques à la production d'un travail technique en langue anglaise: mise en œuvre d'un projet complexe en langue anglaise, communication écrite et orale, acquisition d'un vocabulaire spécialisé, maîtrise de différents niveaux de langues en terme de style (oral vs. écrit) et de technicité (vulgarisation vs. spécialisation), mise en place de stratégies pour faire face à des difficultés linguistiques.

Principales notions abordées :

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

UE Projets

PROJET DE FIN D'ETUDES

Methodological project

Enseignant : Divers intervenants
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)

Répartition des enseignements
Langue d'enseignement : 2,5
: Entre 60h et 75 h

: 9h d'ateliers, et suivis réguliers
: Français
Logiciels : /
Documents pédagogiques : /
Pré-requis : /

Modalités d'évaluation :

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. Le projet est évalué à travers un rapport et une soutenance.

Acquis d'apprentissage (objectifs) :

Les objectifs du projet de fin d'études, et donc les compétences qui sont renforcées grâce à celui-ci, sont multiples:

- mise en situation professionnelle
- capacité à définir une stratégie d'étude en réponse à une demande client
- mobilisation des compétences techniques (statistiques, économiques, informatiques)
- compromis entre rigueur scientifique et contraintes pratiques (limitations financières, logicielles, cognitives, temporelles…)
- travail de groupe
- gestion d'un projet sur le temps long
- communication (écrite, orale) sur des sujets techniques

.

Principales notions abordées :

Travail autonome en groupe suivi par un professionnel de l'entreprise ou de la recherche (env. 5 séances).

UE Projets

DATA CHALLENGE

Data Challenge

Enseignant : Divers intervenants industriels (correspondante : Salima El Kolei)
Nombre d'ECTS
Volume horaire de travail élève
(enseignements + travail personnel)

Répartition des enseignements
Langue d'enseignement : Pas d'attribution d'ECTS
: 2 journées

: 12h d'ateliers
: Français
Logiciels : /
Documents pédagogiques : /
Pré-requis : Méthodes de travail des projets, Compétences statistiques et informatiques de 3ème année

Modalités d'évaluation :

Les élèves participent au data challenge proposé à l'Ensai 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 0

UE : PROJET PROFESSIONNEL ET STAGES

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

: 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'Ensai. 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'Ensai 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

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

Descriptifs des enseignements de la filière

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

UE Spécifiques filière SSV

UE SPÉCIFIQUES FILIÈRE SSV

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

: 257h

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)
- Maitriser 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).

UE2 - Méthodologie statistique 1

Plans d'expériences
Experiment Design

Cours : 18h

Enseignant	Walter TINSSON (Université de PAU)
Correspondant	Myriam VIMOND

Enseignement destiné aux élèves des filières « Statistique pour les sciences de la vie» et « Génie statistique » , et « Ingénierie statistique des territoires et de la santé »

Objectif pédagogique

Comprendre les principes fondateurs des stratégies d'expérimentation
Apprendre à choisir, construire un dispositif expérimental
Acquérir les outils d'analyse des plans d'expériences (utilisation du logiciel R)

Contenu de la matière

Principes fondateurs et présentation des grandes familles de plans
Les Outils d'Analyse : modèle linéaire, analyse de la variance
Plans factoriels complets et fractionnaires, Optimalité
Expériences Accélérées

Pré-requis

Ce qui a été vu en classes préparatoires et lors des deux premières années de l'Ecole est largement suffisant pour suivre le cours (Calcul matriciel, Optimisation, Bases de la régression et d'analyse de variance…)

Contrôle des connaissances

Examen pratique en salle informatique

Références bibliographiques

- AZAIS, J.-M., BARDET, J.-M. Le modèle linéaire par l'exemple (2e éd.). Dunod, 2012.
- DROESBEKE, J.-J., FINE, J., SAPORTA, G. (Eds Scientifiques). Plans d'expériences : Applications à l'entreprise. Technip, 1997.

Langue d'enseignement

Français

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

UE2 - Méthodologie statistique 1

Modèles mixtes
Mixed Models

Cours : 15h • Atelier : 6h

Enseignant : Etienne DANTAN (Université de Nantes)

Correspondant : Myriam VIMOND

Enseignement destiné aux élèves de la filière « Statistique pour les sciences de la vie »

Objectif pédagogique

A l'issue de cet enseignement, les élèves devront connaître les fondements de la théorie statistique du modèle mixte afin d'en assurer une bonne compréhension, maîtrise et interprétation.

Contenu de la matière

Dans le modèle linéaire, on peut prendre en compte diverses structures de corrélation (intra-classe, temporelle, spatiale) grâce à une version dite "mixte" du modèle qui fait intervenir à la fois des effets fixes et des effets aléatoires.
Il est à noter que nombre de problèmes peuvent être abordés sous une formulation de modèle mixte, cf. le filtre de Kalman et les splines cubiques par exemple. Ce type de modèles peut faire l'objet d'une grande variété de traitements et d'interprétation statistique (algorithme EM, méthodes MCMC, approche bayésienne). Le concept permet également de nombreuses extensions (modèle linéaire généralisé, modèle non linéaire).
Le modèle mixte connaît actuellement un développement important de ses applications dans maints secteurs de l'industrie, des sciences économiques et sociales, de la biologie et de la médecine. Il est servi par de bons logiciels tels que les Proc GLM, Mixed, Nlmixed, Glimmix de SAS, Asreml , Nlme de R et S plus, Monolix, Genstat et Winbugs.

1. Introduction: écriture générale, hypothèses, exemples
2. Prédiction des effets aléatoires: meilleure prédiction, meilleure prédiction linéaire et meilleure prédiction linéaire sans biais (BLUP)
3. Equations du modèle mixte d'Henderson
4. Inférence par la méthode du maximum de vraisemblance : estimation et tests d'hypothèse des effets fixes et des composantes de variance
5. Concept de vraisemblance résiduelle (REML) : présentation classique et présentation bayésienne
6. Théorie de l'algorithme EM
7. Application aux composantes de variance
8. Aperçu sur le modèle linéaire généralisé mixte

Pré-requis

Lois de probabilité, algèbre linéaire et théorie du modèle linéaire (regréssion, anova)

Contrôle des connaissances

Examen écrit.

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, 2012G. VERBEKE, G. MOLENBERGHS, Linear mixed models in practice, Springer Verlag, New York, 1997

Langue d'enseignement

Français

UE2 - Méthodologie statistique 1

Compléments de modèles de durée
Survival Analysis Applied to Biostatistics

Cours : 15h • Atelier : 6h

Enseignant : Lisa BELIN (Sorbonne Université- INSERM- AP-HP
David HAJAGE (Sorbonne Université- INSERM- AP-HP)

Correspondant : Myriam VIMOND

Enseignement destiné aux élèves de la filière « Statistique pour les sciences de la vie »

Objectif pédagogique

A l'issue de cet enseignement, les élèves devront maîtriser les connaissances de base pour analyser des données censurées. Ces problèmes concernent tous ceux qui sont impliqués dans les sciences de la vie où ce type de données est très fréquemment rencontré. A l'issue de ce module, l'étudiant doit être capable de mettre en œuvre les méthodes enseignées pour résoudre les problèmes classiques de comparaison de plusieurs échantillons, de modélisations (paramétriques ou non) des distributions de survie.

Contenu de la matière

Les quatre premières séances seront consacrées à l'enseignement théorique des différentes méthodes. Une dernière séance consistera en un apprentissage des différentes procédures disponibles dans les logiciels (SAS en particulier). 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.

1- Généralités et particularités des données de survie.
2-Définition des différentes fonctions de survie.
3-Recueil des données de survie et préparation de la base de données à analyser.
4-Estimations non paramétriques des fonctions de survie.
5-Estimations et comparaisons paramétriques des distributions de survie - Cas du modèle exponentiel. Modèle exponentiel généralisé. Etude graphique de l'adéquation de la modélisation.
6-Comparaisons non paramétriques de plusieurs distributions de survie : logrank pondérés, liens avec les tests de rangs.
7-Calcul du nombre de sujets nécessaires : cas du modèle exponentiel et du logrank.
8-Modélisation semi-paramétrique des fonctions de survie par le modèle de Cox : définition de la vraisemblance conditionnelle, problème de codages, modèles avec interactions, étude de l'adéquation de l'hypothèse des taux proportionnels.
9-Introduction à différents problèmes : variables dépendantes du temps, risques concurrents, recherche d'interactions qualitatives, analyses intermédiaires …
10-Etude d'un cas en vraie grandeur : Recherche de facteurs pronostiques chez des enfants atteints de tumeurs cérébrales.
11-Mise en application (Y. De Rycke).

Pré-requis

Aucun

Contrôle des connaissances

Examen écrit (2h et documents autorisés).

Références bibliographiques

- C. HILL, C. COM-NOUGUE, A. KRAMAR, T. MOREAU, J. O'QUIGLEY, R. SENOUSSI, C. CHASTANG, Analyse statistique des données de survie (2e éd.), Collection statistique en biologie et médecine, Flammarion, 1996.

Langue d'enseignement

Français

UE3 - Méthodologie statistique 2

Mesures de qualité de vie
Measuring Quality of Life

Cours : 9h • Atelier : 6h

Enseignants : Jean Benoît HARDOUIN (Université de Nantes)

Correspondant : Myriam VIMOND

Enseignement destiné aux élèves de la filière « Statistique pour les sciences de la vie »

Objectif pédagogique

Les mesures de qualité de vie associée à la santé (Health-Related Quality of Life Measures) ont connu au cours des 20 dernières années un développement important, notamment dans le domaine de l'évaluation des stratégies thérapeutiques du cancer, du SIDA ou encore de maladies chroniques. Les mesures de qualité de vie font partie des « mesures subjectives en santé », par opposition aux mesures d'efficacité cliniques objectives, traditionnellement utilisées dans les évaluations. Elles font appel à des méthodes et concepts développés en psychométrie, mais aussi en économie de la santé (approche par les préférences individuelles).

Contenu de la matière

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"

Pré-requis

Contrôle des connaissances

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

Langue d'enseignement

Français

UE3 - Méthodologie statistique 2

Statistique des processus
Statistics of Stochastic Processes

Cours : 12h Atelier : 3h

Enseignant : Myriam VIMOND **(**Ensai)

Correspondants : Salima EL KOLEI

Enseignement avec une partie commune aux élèves des filières « Statistique pour les sciences de la vie » et « Génie statistique »

Objectif pédagogique

Il s'agit de présenter des modélisations des principaux phénomènes aléatoires dépendants du temps rencontrés dans l'industrie et en sciences de la vie (hors-séries temporelles et traitement du signal) à partir de processus markoviens. Dans chaque cas la présentation portera autant sur les outils probabilistes que sur l'inférence statistique dans ces modèles.

Contenu de la matière

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
Files d'attente (uniquement pour « Génie Statistique »)

Pré-requis

Les cours de Probabilités et Statistique de première année et le cours de chaînes de Markov

Contrôle des connaissances

Examen écrit.

Références bibliographiques

- ASMUSSEN, S. Applied probability and queues. Second edition. Springer 2003.
- BOSQ, D. Statistique mathématique et statistique des processus. Lavoisier, 2012.
- DACUNHA-CASTELLE, D., DUFLO, M. Probabilités et statistiques II, Problèmes à temps mobile. Masson, 1993.
- DELMAS, J-F., JOURDAIN, B. Modèles aléatoires. Applications aux sciences de l'ingénieur et du vivant. Springer 2006.
- FOATA, D., FUCHS,A. Processus Stochastiques (2e éd.). Dunod 2004.
- FUCHS, C., Inference for Diffusion Processes, With Application in Life Sciences. Springer 2013.
- KIMMEL, M., AXELROD, D. Branching processes in biology. Springer 2002.
- PARDOUX, E. Processus de Markov et applications. Algorithmes, génome et finance. Dunod 2007.

Langue d'enseignement

Français

UE3 - Méthodologie statistique 2

Traitement des données manquantes dans les essais cliniques
Handling of missing data in clinical trials

Cours : 6h • Atelier : 6h

Enseignant : Mélanie PRAGUE (INRIA)

Correspondant : Myriam VIMOND

Enseignement destiné aux élèves de la filière « Statistique pour les sciences de la vie »

Contenu de la matière

PART I Introduction to missing data – issues with missing data
Assumptions about missing data and missing data mechanisms – A formal taxonomy: MCAR, MAR and MNAR
Regulatory considerations – The FDA-commissioned NRC/NAS report – Notion of estimand
Single imputation methods: LOCF, BOCF
Commonly used analytic methods under MAR:

- Complete-Case-Analysis (CCA)
- A modification of CCA under MAR: Inverse Probability Weighting (IPW)

PART II Commonly used analytic methods under MAR (continued):

- Single imputation
- Multiple imputation
    - The multivariate normal regression framework
    - Monotone missing data patterns
    - Non-monotone missing data patterns: MCMC algorithms
    - Case of binary outcomes

PART III Parametric analyses under MAR:

- Likelihood-based methods (Mixed Model with Repeated Measures = MMRM)
- Case study – Computer practical using SAS (version 9.4 or higher) – Questions 1-5
- Marginal versus conditional models

Introducing MNAR assumptions via the conditional model:

- Shift parameter
- Power considerations

Pattern-Mixture-Models (PMM)

PART IV Principles and methods of sensitivity analyses under MNAR via PMM:

- Control-based imputation: Copy Reference (CR) and Jump To Reference (J2R) inference
- Case study – Computer practical using SAS – Question 6
- Delta adjustment and tipping point analysis
- Case study – Computer practical using SAS – Question 7
- Power considerations for MMRM and delta-adjusted PMM analyses
- Case study – Computer practical using SAS – Question 8

Contrôle des connaissances
A déterminer

Langue d'enseignement

Français mais supports écrits en anglais

UE3 - Méthodologie statistique 2

Statistique Bayésienne
Advanced Bayesian Statistics

Cours : 9h • Atelier : 6h

Enseignant : Sophie ANCELET (IRSN)

Correspondant : Myriam VIMOND

Enseignement destiné aux élèves de la filière "Génie Statistique", « Ingénierie statistique des territoires et de la santé » et « Statistique pour les sciences de la vie »

Objectif pédagogique

A l'issue de cet enseignement, les élèves devront maîtriser les connaissances de base pour l'analyse de données par approche statistique bayésienne. Les problèmes traités seront empreints aux sciences de la vie où l'emploi des méthodes bayésiennes progresse considérablement. Cependant, le caractère général de ces méthodes, utiles dans bien d'autres domaines d'application, sera clairement établi. À l'issue de ce module, l'étudiant doit être capable de mettre en œuvre les méthodes enseignées pour mener des inférences bayésiennes de données, notamment à l'aide des logiciels WINBUGS, OPENBUGS et JAGS.

Contenu de la matière

Un rappel de cours est fait concernant les principes de la modélisation statistique bayésienne. L'accent sera mis sur l'analyse bayésienne par les méthodes de Monte Carlo par Chaînes de Markov (MCMC). Aux travers d'exemples, seront abordés les notions de graphe d'indépendance conditionnelle, réseau bayésien, convergence des Chaînes de Markov, inférence, prédiction, validation et comparaison de modèles dans un cadre bayésien. Les exemples seront traîtés sous le logiciel WINBUGS ou JAGS en salle informatique.

Pré-requis

Cours de statistique bayésienne en deuxième année

Contrôle des connaissances

Projet court

Références bibliographiques

- Collectif BIOBAYES: Albert I., **Ancelet S.**, David O., Denis J.B., Makowski D., Parent E., Soubeyrand S. **(2015)** Méthodes statistiques bayésiennes. Bases théoriques et applications en alimentation, environnement et génétique**.** *ELLIPSES, ISBN : 978234000501*
- Carlin, B. P. and Louis, T.A. (2009). Bayesian Methods for Data Analysis. Chapman & HALL/CRC, third edition, (535 pp.)
- Gelman, A., Carlin, J. B., Stern, H. S. and. Rubin, D. B (2004). Bayesian data analysis. Texts in Statistical Science. Chapman & HALL/CRC, second edition, (668 pp.)
- Robert, C. P. (2001). The Bayesian choice. Springer, (second edition) (604 pp.)
- Lunn, D.J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing, 10: 325-337.
- Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). Markov chain Monte Carlo in practice. Chapman and Hall, (486 pp.)

Langue d'enseignement

Français

Campus de Ker Lann, 51 rue Blaise Pascal - BP37203 - 35172 BRUZ CEDEX
Tél : 33 (0)2 99 05 32 32 / scolarite@ensai.fr

www.ensai.fr

UE3 - Méthodologie statistique 2

Méta-Analyse
Meta Analysis

Cours : 12h • Atelier : 6h

Enseignant : Drifa BELHADI (Creativ Ceutical)

Correspondant : Samuel DANTHINE

Enseignement destiné aux élèves des filières « Ingénierie Statistique des Territoires et de la Santé» et « Statistique pour les sciences de la vie »

Objectif pédagogique

"La méta-analyse est une démarche, plus qu'une simple technique, qui a pour but de combiner les résultats de plusieurs essais thérapeutiques, pour en faire une synthèse reproductible et quantifiée. Cette synthèse produit un gain de puissance statistique dans la recherche de l'effet d'un traitement, une précision optimale dans l'estimation de la taille de l'effet et permet en cas de résultats apparemment discordants d'obtenir une vue globale de la situation".
Trois types de méta-analyses sont distingués, en fonction des données utilisées:

1. La méta-analyse des données résumées de la littérature, donc uniquement des essais publiés (ce qui expose au biais de publication)
2. La méta-analyse exhaustive sur données résumées se basant sur les études publiées et sur les travaux non publiés
3. La méta-analyse sur données individuelles se basant sur les données de tous les patients inclus dans les essais pris en considération dans la méta-analyse.

Dans la démarche de la méta-analyse, la variabilité (l'hétérogénéité) est considérée comme un paramètre de nuisance; elle contredit l'hypothèse de l'existence d'un effet traitement commun à tous les essais. La méta-analyse est très utilisée, notamment dans les analyses médico-économiques qui utilisent dans leur modélisation des indicateurs de résultats issus de publications diverses.

Contenu de la matière

1. Introduction
   1.1. The use of meta-analysis in clinical trial / Health economic evaluation
   1.2. Meta-analysis vs. randomised clinical trials
2. Protocol development
   2.1. Outcome measure and baseline information
   2.2. Data sources / Study selection
   2.3. Data extraction
   2.4. Analyses / Sensitivity analyses
   2.5. Presentation of results
3. Estimating treatment difference
   3.1. Binary data (Log-odds ratio and Log-relative risk)
   3.2. Normally distributed data (Absolute mean difference and Standardised mean difference)
   3.3. Ordinal data (Log-odds ratio (prortional odds model))
   3.4. Survival data (Log hazard ratio)
4. Combining estimates of treatment difference
   4.1. Fixed-effects parametric approach (FE)
      4.1.1. Estimation of the treatment difference and hypothesis test
      4.1.2. Testing for heterogeneity

4.2. Random-effect parametric approach (RE)
4.2.1. Estimation of the treatment difference and hypothesis test
4.2.2. Testing for between studies heterogeneity
5. Dealing with heterogeneity
5.1. Limited power of heterogeneity tests
5.2. Choice between FE and RE models
5.3. Can we always present an overall estimate of treatment difference?
5.4. Choice of appropriate measure of treatment difference
5.5. Meta regression
6. Presentation of results
7. Selection / publication bias
8. Direct comparison vs. Indirect comparison
8.1. Eg. Drug A vs placebo & Drub B vs. placebo => Drug A vs. Drug B
9. An introduction to Bayesian approach
10. Conclusion
10.1. The use of meta-analysis
10.2. Contrast between useful and useless meta-analysis

Pré-requis

Basic Winbugs knowledge

Contrôle des connaissances

A déterminer

Références bibliographiques

Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions. Chichester (UK): John Wiley & Sons, 2008.

Dias, S., Welton, N.J., Sutton, A.J. & Ades, A.E. NICE DSU Technical Support Document 2: A Generalised Linear Modelling Framework for Pairwise and Network Meta-Analysis of Randomised Controlled Trials. 2011; last updated September 2016; available from http://www.nicedsu.org.uk

Langue d'enseignement

Anglais

UE4 - Essais cliniques

Essais cliniques : méthodologie et analyses statistiques
Clinical Trials

Cours : 18h

Enseignants : David HAJAGE (Sorbonne Université- INSERM- AP-HP), Lisa BELIN (Sorbonne Université- INSERM- AP-HP) et Yann DE RYCKE

Correspondant : Myriam VIMOND

Enseignement destiné aux élèves des filières « Statistique pour les sciences de la vie » et « Ingénierie des territoires et de la santé »

Objectif pédagogique

La nature et la structure des données recueillies dans le cadre d'essais cliniques (qui incluent les études sur les médicaments, les interventions médicales novatrices et les nouveaux matériels) nécessitent de recourir à des méthodes statistiques adaptées.
Cet enseignement permettra aux élèves de se familiariser avec les différents types d'études, les enjeux, les acteurs et plus particulièrement les méthodes statistiques utilisées dans le domaine des essais cliniques.

Contenu de la matière

Après une présentation générale des essais cliniques, le cours comportera 2 parties. La première aura comme objectif de permettre aux élèves de se familiariser avec la méthodologie des essais cliniques et de découvrir le déroulement d'une étude du point de vue du biostatisticien. La seconde s'attachera à détailler certaines méthodes utilisées dans l'analyse des études cliniques.

Présentation générale des essais cliniques :
Les différentes étapes d'une étude, les intervenants, le rôle du biostatisticien
Aspects règlementaires et éthiques
Déroulement d'une étude pour le biostatisticien
L'analyse statistique : du plan d'analyse aux résultats
Choix d'une méthode adaptée aux objectifs et aux données
Puissance et nombre de sujets nécessaires
Rédaction d'un rapport statistique
Divers éléments à prendre en considération : Biais, Indépendance, Normalité, Bilatéral / Unilatéral, Populations ITT et PP, …
Approfondissement de quelques méthodes d'analyse
Mesures répétées
Essais de différence, d'équivalence, de supériorité, de non infériorité
Courbes ROC
Concordance, fiabilité, reproductibilité

Contrôle des connaissances

Examen écrit

Références bibliographiques

Sera distribuée en séance

Langue d'enseignement

Français

UE4 - Essais cliniques

Pharmacométrie
Pharmacometrics

Cours : 6h • Atelier : 12h

Enseignants	:	Jérémie GUEDJ, Emmanuelle COMETS (INSERM, Institut Claude Bernard University Paris Diderot)
Correspondant	:	Myriam VIMOND

Enseignement destiné aux élèves de la filière « Statistique pour les sciences de la vie »

Objectif pédagogique

A l'issue de cet enseignement, les élèves devront maîtriser les principales méthodes statistiques utilisées par le biostatisticien lors de l'analyse ou la conception des essais cliniques.

Contenu de la matière

Introduction à la pharmacométrie, la pharmacocinétique et la pharmacodynamie (principes, rôle dans le développement des médicaments, exemples de modèles)
Modèles non-linéaires à effets mixtes (introduction, historique, modèles statistiques, méthodes d'estimation, logiciels)
Construction et évaluation de modèles
TP en Monolix (présentation du logiciel, construction d'un modèle pharmacocinétique)

Optimisation de protocoles dans les modèles non linéaires simples ou mixtes, théorie et applications, simulation d'essais cliniques
TP en Monolix et PFIM (graphes diagnostiques, simulation, optimiation)
Séminaire

Pré-requis

Modèles Mixtes

Contrôle des connaissances

Examen

Références bibliographiques

Langue d'enseignement

Français

UE4 - Essais cliniques

Projet d'essais cliniques
Project in Clinical Trials

Atelier : 6h • Projet : 18h

Enseignant : Caroline PETIT (Sanofi)

Correspondant : Myriam VIMOND

Enseignement destiné aux élèves de la filière « Statistique pour les sciences de la vie »

Objectif pédagogique

Le but de ces projets est de mettre en application quelques-unes des méthodes vues pendant le cours sur les essais cliniques. Dans cet objectif, une base de données correspondant à un essai clinique réel est fournie aux élèves, ainsi que les documents nécessaires (cahier d'observation, protocole...) afin qu'ils réalisent suivant des directives précises l'analyse demandée, en utilisant le logiciel R.

Contenu de la matière

Ce projet, réalisé en général par groupe de 2 ou 3 élèves, 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.
A la fin du projet, les élèves doivent remettre un rapport statistique d'une trentaine 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. Les élèves auront à ce titre un soutien assuré par un enseignant d'anglais.

Pré-requis

Introduction aux essais cliniques, compléments d'essais cliniques. Bonne maîtrise des méthodes statistiques générales, en particulier: statistiques descriptives, régression linéaire, analyse de variance, régression logistique, analyse des données de survie, modèles mixtes.

Contrôle des connaissances

L'évaluation des projets repose à la fois sur le contenu des rapports et le comportement à la soutenance, devant un jury composé de spécialistes.

Références bibliographiques

Sera distribuée en séance

Langue d'enseignement

Français. Le rapport sera rédigé en anglais, et donnera lieu à une soutenance en anglais.

UE5 – Epidémiologie - Génomique

Epidémiologie quantitative
Quantitative Epidemiology

Cours : 15h

Enseignant : Olivier GRIMAUD , Pascal CREPEY, Cindy PADILLA (EHESP)

Correspondant : Myriam VIMOND

Enseignement destiné aux élèves de la filière « Statistique pour les sciences de la vie »

Objectif pédagogique

L'objet de l'épidémiologie est l'étude la distribution dans le temps et dans l'espace des états de santé des populations humaines et l'analyse leurs déterminants.
A l'issue du cours, l'étudiant sera capable de :
- Procéder à une description simple de la situation épidémiologique à partir des indicateurs épidémiologiques usuels.
- Formuler les hypothèses de liaison, discuter de la possibilité d'une relation de cause à effets entre les facteurs de risque et l'état de santé.
- Choisir les types d'études appropriés à la mise en évidence des relations entre maladies et facteurs de risque.
- Interpréter les résultats d'études épidémiologiques.

Contenu de la matière

Intitulé de la séance de cours
Introduction de l'épidémiologie – Causalité – Enquêtes de cohorte
Epidémiologie descriptive: prévalence, incidence, standardisation
Les mesures d'association
Enquêtes cas-témoins. Biais
Facteurs de confusion
Analyses multivariées

Pré-requis

Les méthodes statistiques de base sont supposées connues (fluctuations d'échantillonnage, intervalle de confiance, méthodes d'estimation et tests statistiques usuels).

Contrôle des connaissances

L'évaluation consistera en un examen écrit de 1 heure et demi, tous documents et calculette sont autorisés.

Références bibliographiques

- J. BOUYER, D. HEMON, S. CORDIER, F. DERRIENNIC, I. STUCKER, B. STENGEL, J. CLAVEL, Epidémiologie. Principes et méthodes quantitatives, Paris, Les éditions INSERM, 1993

- K.J. ROTHMAN, S. GREENLAND, Modern epidemiology, Little, Brown and Company, Boston, 1998
- D.G. KLEI NBAUM, L.L. KUPPER, H. MORGENSTERN, Epidemiologic Research. Principles and quantitative methods, New York, Van Nostrand Reinhold, 1982

Langue d'enseignement

Français

UE5 – Epidémiologie - Génomique

Modélisation compartimentale
Compartmental Modeling

Cours : 6h • Atelier : 6h

Enseignant	Audrey LAVENU (Université de Rennes 1)
Correspondant	Myriam VIMOND

Enseignement destiné aux élèves de la filière « Statistique pour les sciences de la vie »

Objectif pédagogique

A l'issue de cet enseignement, les élèves devront être capables de simuler des épidémies par des modèles Susceptible-Infectious-Removed (SIR), de comprendre l'interprétation des paramètres et de construire des modèles dérivés du modèle standard.

Contenu de la matière

Introduction : Contexte de la modélisation compartimental. Définition d'une épidémie.
1. Modèle SIR déterministe
1.1. Construction du modèle (système d'équations différentielles)
1.2. Calcul du taux de reproduction de base (R0)
1.3. Simulation d'épidémies
1.4. Exemples de données de surveillance épidémiologique
- le réseau Sentinelles
- application sur logiciel

1.5. Exemples de modèles dérivés du modèle SIR.
2. Modèle SIR stochastique

Pré-requis

Aucun

Contrôle des connaissances

Rapport basé sur exercices commencés en séance avec commentaires étoffés du cours

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.
- Bailey N. T. J., 1975: The mathematical theory of infectious diseases and its application. Griffin, London, 2nd edition (épuisé).
- Bartlett M. S. 1960. Stochastic Population Models in Ecology and Epidemiology,
- Methuen, London.

Langue d'enseignement

Français

UE5 - Epidémiologie - Génomique

Analyse des données « Omiques »
Omics Data Analysis

Cours : 6h • Atelier : 12h

Enseignants : Isabelle BRITO (Institut Curie) et Pierre NEUVIAL (IMT)

Correspondant : Myriam VIMOND

Enseignement destiné aux élèves de la filière « Statistique pour les sciences de la vie »

Objectif pédagogique

Ces séminaires permettent aux étudiants de rencontrer différents chercheurs travaillant sur l'analyse d'une grande variété de données « omics », y compris les données génomiques, transcriptomiques, métabolomiques, protéomiques, épigénétiques, et métagénomiques.

Contenu de la matière

Introduction à la génomique.
Présentation des données en épidémiologie génétique, et des données omiques en recherche clinique.
Méthodes de recherches de classes, indices de validation de classes.
Mesures d'association allélique : le déséquilibre de liaison (LD), l'équilibre de Hardy-Weinberg (HWE))
Méthodes de comparaisons de classes.
Tests multiples : Family-Wise Error Rate (FWER), False Discovery Rate (FDR)
Méthodes de sélection de variables.

Pré-requis

Introduction à l'analyse de données « Omics », théorie des tests, modèle linéaire (ANOVA et modèle mixte), classification.

Contrôle des connaissances

Un examen écrit qui comprend les notions abordées en cours et des extraits de scripts R.

Références bibliographiques

- Mary-Huard T., Picard F., Robin S. Introduction to Statistical Methods for Microarray Data Analysis, in Mathematical and Computational Methods in Biology, Hermann : Paris, 2007.
- Gentleman R.C., Carey V.J., Dudoit S., Irizarry R., Huber W., Bioinformatics and Computational Biology Solutions using R and Bioconductor, New York: Springer, 397-420, 2005.

Langue d'enseignement

Français

UE5 - Epidémiologie - Génomique

Introduction à l'analyse de données « Omiques »
Introduction to Omics Data Analysis

Cours : 6h • Atelier : 6h

Enseignant : Guillaume DESACHY (AstraZeneca)

Correspondant : Myriam VIMOND

Enseignement destiné aux élèves de la filière « Statistique pour les sciences de la vie »

Objectif pédagogique

Les séminaires sur le génome permettent aux étudiants de rencontrer différents chercheurs travaillant sur des problèmes liés aux génomes.

Contenu de la matière

L'étude de la génomique a de nombreuses applications dans le domaine des sciences de la vie. Les problèmes liés au génome font appel à de nombreuses méthodes statistiques. Différentes approches pour l'analyse des séquences génomiques seront présentées.
Le cours consiste en une vaste initiation à la Génomique et Post-Génomique (Transcriptomique, Protéomique) et de façon plus générale à la Bioinformatique. Les objectifs de ce cours sont l'acquisition des connaissances de bases (du Génome à l'Organisme en passant par l'ADN, l'ARN, la protéine, la cellule et les bases de données associées) permettant la compréhension des principales problématiques liées à l'application de la Statistique à la Biologie. Les principales méthodes statistiques évoquées sont les chaînes-de-markov, chaînes-de-markov cachées, test d'hypothèse, test-multiple, la classification et l'alignement de séquences biologiques. Le cours est accompagné de travaux pratiques réalisés sous R.

Pré-requis

Pas de pré-requis

Contrôle des connaissances

Mini-projets en anglais, compte-rendus de TP.

Langue d'enseignement

Français