université PARIS-SACLAY

ÉCOLE D'ÉTÉ DE SANTÉ PUBLIQUE ET D'ÉPIDÉMIOLOGIE DE BICÊTRE

Institut National de la Santé et de la Recherche Médicale
Faculté de Médecine Paris-Saclay
Ecole de Santé Publique Paris-Saclay
Responsable : Laurence Meyer

Du 23 juin au 4 juillet 2025

Contact : Sandra Clares Vera - Tél. (+33)01.49.59.18.54 - E-mail : ecole.ete@inserm.fr

Ecole d'été de Santé Publique et d'Epidémiologie de Bicêtre - Inserm CESP U1018
82, rue du Général Leclerc - 94276 LE KREMLIN-BICÊTRE cedex, France

Plus d'informations : https://eespe.vjf.inserm.fr

L'école d'Eté de Santé Publique et d'Epidémiologie de Bicêtre comprend des enseignements théoriques dans les domaines de la biostatistique, de l'épidémiologie et de l'informatique et des enseignements pratiques concernant plus particulièrement la formulation, la résolution et la discussion de problèmes concrets de Santé Publique. L'accent est surtout mis sur l'utilisation des connaissances acquises au cours de cet enseignement. Les enseignants exercent une activité de recherche dans une Unité de recherche de l'INSERM, dans un laboratoire universitaire ou dans une Agence de Santé. La participation aux enseignements donne droit à la délivrance d'une attestation.

D'avance, merci pour votre participation !

Laurence Meyer, responsable de l'école d'été d'Epidémiologie et de Santé Publique de Bicêtre

---

Personnes concernées

Le programme propose un enseignement de base (1) et des enseignements approfondis (2 à 7) en méthodologie et applications pour répondre aux besoins des professionnels de la Santé, décideurs, médecins, chercheurs, qui souhaitent découvrir ou approfondir les bases du raisonnement en sciences de la Santé et son utilisation dans le domaine de la Santé Publique. L'enseignement de base concerne plus particulièrement les personnes qui n'ont aucune formation en statistique et en épidémiologie. Tout dossier devra, pour être accepté, avoir été examiné par le Comité de Direction de l'Ecole d'Eté. Le nombre de participants est limité.

Modalités et droits d'inscription

Le formulaire de pré-inscription devra être rempli en ligne (https://eespe.vjf.inserm.fr) ou retourné avant le 6 juin 2025 accompagné d'une photo d'identité. Le règlement du montant total des frais d'inscription est obligatoire pour toute pré- inscription à titre individuel. Le montant total de ces frais figure ci-après :

- 660 euros pour un enseignement à temps complet (sur deux semaines)
- 330 euros pour un enseignement à mi-temps sur deux semaines, ou à temps plein sur une semaine (modules 2, 3, 7)
- 165 euros pour un enseignement à temps plein sur deux jours et demi ou mi-temps sur une semaine (modules 4, 5, 6)
- 2100 euros dans le cadre de la Formation Permanente ou pour les entreprises présentant un candidat (1100 euros si enseignement à mi-temps).

Pour les inscriptions au titre de la Formation Permanente, consulter le site internet ou contacter le secrétariat. Les personnes s'inscrivant dans ce cadre devront faire établir par leur employeur une lettre de prise en charge. Une Convention de Formation Professionnelle sera alors établie par la Faculté de Médecine Paris-Saclay.

Après la notification de l'acceptation de votre dossier, un courrier vous sera adressé comprenant les indications pratiques. Les documents pédagogiques seront transmis aux participants via le site de l'école d'été.

Liste des modules pour 2025

Module 1 : Méthodologie de base en statistique et en épidémiologie

F. Boufassa - 2 semaines du 23 juin au 4 juillet, temps complet

Module 2 : Régression multiple en épidémiologie : régression linéaire, logistique et modèle de Cox

J.-P. Empana et B. Heude - 1 semaine du 23 au 27 juin, temps complet

Module 3 : Introduction à la modélisation des données longitudinales sous R : modèles marginaux et modèles mixtes

H. Panjo et F. Artaud - 1 semaine du 30 juin au 4 juillet, temps complet

Module 4 : Stratégie de recherche en épidémiologie

M. Canonico - 2 jours et demi du 23 au 25 juin, temps complet

Module 5 : Initiation à l'analyse de données avec R (tidyverse)

R. Bauer et S. Guitton - 1 semaine du 23 au 27 juin, après-midi

Module 6 : R avancé, automatiser ses rapports d'analyses avec R

R. Bauer et S. Guitton - 1 semaine du 30 juin au 4 juillet, après-midi

Module 7 : Introduction au *machine learning* en recherche biomédicale

M. Sedki et J.-P. Teglas - 1 semaine du 23 au 27 juin, temps complet

1. Méthodologie de base en statistique et en épidémiologie

F. Boufassa

Objectifs

Ce module a pour but de fournir aux professionnels et responsables de la Santé des notions de base en statistique et en épidémiologie.

Modalités Pédagogiques

Les cours théoriques seront enrichis d'exemples pratiques ou de sorties de listing de logiciels d'analyse explicités.

Pré-requis

Ce module s'adresse aux personnes qui n'ont aucune formation en statistique et en épidémiologie.

Programme

- Méthodologie statistique : introduction à la notion de variabilité, statistique descriptive, principe des tests
- Introduction à l'épidémiologie : différents types d'enquêtes, organisation d'une enquête
- Mesures de fréquence : taux, standardisation, ...
- Evaluation des méthodes diagnostiques : sensibilité, spécificité et valeurs prédictives
- Notions de risque, biais et facteurs de confusion en épidémiologie
- Puissance d'une enquête, nombre de sujets nécessaire
- Analyse critique d'un article en épidémiologie
- La méthode d'ajustement de Mantel - Haenszel

2. Régression multiple en épidémiologie : régression linéaire, logistique et modèle de Cox

J.-P. Empana et B. Heude

Objectifs et contenu

L'objectif du cours est de présenter les trois principaux modèles de régression multivariée les plus fréquemment utilisés en épidémiologie et en recherche biomédicale : la régression linéaire, la régression logistique et le modèle de Cox.

Il s'agira de développer une démarche d'analyse, depuis la vérification des hypothèses inhérentes à chaque modèle, le choix et le codage des facteurs de confusion, jusqu'à l'interprétation des résultats. Certains aspects moins usuels seront abordés (régression polytomique dans la régression logistique ou variable dépendante du temps dans le modèle de Cox).

A côté des bases théoriques sous-tendant l'utilisation de ces modèles, de nombreux exemples et situations concrètes seront présentés, notamment sur la base de lectures critiques d'articles. A l'issue de cet enseignement, les étudiants auront acquis les connaissances leur permettant d'utiliser les principaux modèles de régression en fonction de leur problématique, et d'avoir un regard critique dans l'interprétation des résultats.

Pré-requis

Enseignement de base ou cours de statistique et d'épidémiologie de niveau CESAM, M1 de sciences biomédicales, de santé publique ou de mathématiques. Avoir des connaissances de base concernant le modèle de régression linéaire.

Plan

I. Régression linéaire et logistique (5 demi-journées) :

- Rappels sur la régression linéaire
- Régression logistique binaire
    - Présentation du modèle, estimation et tests des paramètres
    - Codage des variables et interprétation des coefficients
    - Evaluation du modèle
    - Sélection des variables selon l'objectif de la modélisation (étiologique ou prédictif)
    - Introduction aux méthodes de sélection de variables par pénalisation (LASSO)
    - Régression logistique polytomique

II. Données de survie (5 demi-journées) :

Problématique des données de survie

Construction des courbes de survie par la méthode de Kaplan-Meier et actuarielle

Comparaison des courbes de survie (test du logrank)

Le modèle de Cox :

- vérification des hypothèses du modèle
- analyse multivariée : procédure manuelle vs. procédure automatique, étude d'interaction
- variables dépendantes du temps.

Application des modèles de Cox à l'estimation du risque cardiovasculaire individuel (médecine personnalisée).

3. Introduction à la modélisation des données longitudinales sous R : modèles marginaux et modèles mixtes

H. Panjo et F. Artaud

Objectifs

Dans les études longitudinales où l'on mesure pour différents individus à des échéances successives une réponse biologique ou clinique, il est possible de modéliser l'évolution de cette réponse au cours du temps ainsi que l'influence des caractéristiques des sujets sur cette évolution. Cependant il existe une corrélation entre les mesures d'un même sujet qui devra être prise en compte dans l'analyse statistique.

L'objectif de ce module est de présenter, d'une façon pratique et appliquée, les méthodes et les modèles statistiques les plus utilisés pour analyser ce type de données : d'une part, les modèles mixtes et d'autre part, les modèle GEE (Generalized Estimating Equations).

Pré-requis

Ce cours nécessite une bonne connaissance théorique et pratique niveau M2 santé publique des modèles de régression usuels (les modèles de régression linéaire et logistique).

Programme

- Caractéristiques des données longitudinales et implications
- Modélisation des évolutions moyennes (données continues et discrètes)
- Utilisations des modèles à effets mixtes
- Utilisations des modèles marginaux et GEE

Modalités pratiques

Les applications seront réalisées sur les logiciels STATA et R. Les codes de programmation et les sorties seront fournis pour être décrits et interprétées pendant le cours.

4. Stratégie de recherche en épidémiologie

M. Canonico

Objectifs

Décrire et comparer les différentes stratégies de recherche qui peuvent être utilisées dans l'approche d'un problème de santé en population (étude cas-témoins, cohorte, essai d'intervention).

Cet enseignement interactif résumera les points clés d'une étude épidémiologique en prenant comme exemple l'évaluation des effets secondaires d'un traitement. L'apport de la pharmacoépidémiologie dans l'évolution des pratiques médicales sera également discuté.

Pré-requis

Connaissances de base en méthodologie statistique et épidémiologie.

Programme

Stratégies de recherche en population et élaboration d'un protocole.
Bonnes pratiques en recherche épidémiologique.
Régression logistique conditionnelle et modèle de Cox avec exposition dépendante du temps (applications pratiques à partir de sorties SAS).
Interprétation des résultats (biais, puissance, analyse quantitative des données de la littérature, ...) et écriture scientifique.
Impact de données nouvelles sur les pratiques médicales (agences sanitaires, médias,..).

Séance 1 : Stratégies de recherche en population
Elaboration d'un protocole I – Approche Cas-Témoins

Séance 2 : Analyse d'une étude cas-témoin

Interprétation d'une étude cas-témoins

Séance 3 : Elaboration d'un protocole II – Etude de cohorte (protocole et modèle de Cox avec exposition dépendante du temps)

Comparaison cohorte/cas-témoins

Séance 4 : Principe des méta-analyses et applications - interaction/ajustement avant les méta-analyses

Séance 5 : Elaboration d'un protocole III – Réaliser une étude de cohorte de A à Z

Enseignement sur 2 jours et demi : les cours auront lieu du lundi matin au mercredi midi.

5. Initiation à l'analyse de données avec R (introduction au tidyverse)

R. Bauer et S. Guitton

Objectifs

Etre autonome avec le logiciel R pour l'importation d'un tableau de données, la génération de variables, la réalisation de statistiques descriptives, et des tests et modélisations les plus courants en épidémiologie et recherche clinique.

Le but de ce module n'est pas de discuter des résultats d'analyses statistiques mais d'apprendre à utiliser les commandes pour réaliser ces analyses.

Public concerné

Le cours est prévu pour les « vrais débutants » en programmation R. En revanche, des bases théoriques sont requises : théorie des tests, principaux tests univariés, régression linéaire

Pré-requis

Connaissances théoriques en Biostatistique et Epidémiologie :

- au minimum Licence pro en statistique ou M1 de Santé Publique (CESAM ou équivalent)
- idéalement M2 de Santé Publique - Epidémiologie ou Recherche clinique

Connaissances pratiques :

- environnement Windows : pratique courante indispensable.

Programme

- Prise en main du logiciel R et de l'environnement RStudio.
- Manipulation d'une base de données : Importation d'une table, tri sur une variable, sélection de lignes selon des conditions définies, fusion de plusieurs tables, génération de nouvelles variables, etc.
- Production de statistiques descriptives simples et de graphiques.
- La manipulation de base de données et la production de graphiques seront présentées avec l'extension *tidyverse*. L'extension *tidyverse* permet d'effectuer un grand nombre d'opération courantes et vise à faciliter le travail de nettoyage et de préparation de bases de données, préalable à l'analyse statistiques.

Modalités pédagogiques

Cours en salle d'informatique.

Un stagiaire par ordinateur.

Possibilité d'utiliser son ordinateur personnel, avec le logiciel R et l'environnement RStudio préalablement installés (installation gratuite).

6. R avancé, automatiser ses rapports d'analyses avec R

R. Bauer et S. Guitton

Objectifs

Le but de ce module n'est pas de discuter des résultats des analyses statistiques mais d'apprendre à utiliser les commandes pour réaliser ces analyses.

Public concerné

Utilisateur de R souhaitant améliorer l'esthétique de leur rapport de résultats et gagner du temps en automatisant la réalisation de certaines tâches et la production de documents scientifiques.

Pré-requis

Connaissance et utilisation préalable de R.

Programme

- Production de tableaux « prêts à être publiés » avec l'extension *gtsummary*.
- Générer un document word, pdf ou html avec Rmarkdown.
- Initiation à la programmation (boucles, fonction apply, lapply, sapply, création de fonctions).

Modalités pédagogiques

Cours en salle d'informatique.
Un stagiaire par ordinateur.
Possibilité d'utiliser son ordinateur personnel, avec le logiciel R et l'environnement RStudio préalablement installés (installation gratuite).

7. Introduction au machine learning en recherche biomédicale

M. Sedki et J.-P. Teglas

Objectifs

L'objectif du cours est d'apporter une introduction par l'exemple aux principes et aux outils du Machine Learning en recherche biomédicale. Nous allons aborder la problématique dite d'apprentissage supervisé qui vise à prédire une variable cible à partir d'une ou plusieurs autres variables dites explicatives. Nous distinguerons deux cas pour la variable à prédire. L'un est le problème de la régression, où la variable à prédire est quantitative telle que le coût d'un traitement, un taux de mortalité par ville, la concentration en rétinol plasmatique ou la pression artérielle. Dans le problème de la classification, la variable à prédire prend un nombre fini de valeurs ou modalités telles que "survécu" ou "mort" ou le type de cancer d'un échantillon de tissu.

L'objectif des méthodes rassemblées sous le nom Machine Learning (apprentissage statistique en français) est de construire un modèle à partir de données observées, dites données d'apprentissage, qui sera utilisé pour prédire la partie réponse de cas dits tests pour lesquels nous n'observons que les variables explicatives. Il est nécessaire de prédire avec précision les cas tests, mais aussi comprendre quelles variables explicatives affectent le résultat et comment, et aussi d'évaluer la qualité des prédictions.

Il est facile d'appliquer un algorithme pour répondre à l'objectif de prédiction et de nos jours, on peut simplement lancer un logiciel, néanmoins il est important et parfois difficile de comprendre à quel point la méthode fonctionne réellement. Nous essayerons de détailler le fonctionnement d'un ensemble de familles de modèles à travers une série d'exemples d'application sur jeux de données réelles avec le logiciel R.

Pré-requis

- Base en probabilité et statistique.
- Une pratique régulière de la programmation avec R est indispensable.

Programme

- Formalisme de la régression et de la classification.
- Problématique du choix du meilleur modèle dans une famille de modèles et validation croisée.
- Arbres de décision : arbres de régression et classification.
- Méthodes d'agrégation d'arbres : forêts aléatoires et gradient boosting.
- Réseaux de neurones pour la classification d'images : introduction et illustration sur un exemple
- Si le temps le permet : présentation d'une architecture neuronale de type LLM.

Déroulement

Chaque partie du programme sera accompagnée d'un TP R sur un jeu de données réel issu de la recherche biomédicale. Les deux dernières parties sont uniquement consacrées à la présentation des architectures neuronales avec illustration sur des exemples calculés par le formateur en amont sur des ressources informatiques adaptées (GPU et serveur).

Enseignants et membres du comité de direction

ARTAUD F.	INSERM CESP U1018 – EQ04 - Exposome et hérédité
BAUER R.	INSERM SC10-US19 – Villejuif
BOUFASSA F.	INSERM CESP U1018 - EQ05 - Epidémiologie clinique
BRIAND N.	AP-HP URC Necker-Cochin
CANONICO M.	INSERM CESP U1018 - EQ04 - Exposome et hérédité
CLARES VERA S.	INSERM CESP U1018 - EQ05 - Epidémiologie clinique
EMPANA J.P.	INSERM U970 - Paris - Centre de recherche Cardiovasculaire (PARCC)
GUITTON S.	INSERM SC10-US19 - Villejuif
HEUDE B.	INSERM CRESS U1153 - ORCHAD
LAMBERT O.	INSERM CESP U1018 – EQ05 – Epidémiologie clinique
MADEC Y.	Institut Pasteur - Unité d'Epidémiologie des Maladies Émergentes
MAYEN CHACON A.-L.	INSERM CESP U1018 - EQ05 - Epidémiologie clinique
MEYER L.	INSERM CESP U1018 - EQ05 - Epidémiologie clinique
NOVELLI S.	INSERM CESP U1018 - EQ05 - Epidémiologie clinique
PANJO H.	INSERM CESP U1018 - EQ10 - Soins primaires et prévention
SEDKI M.	INSERM CESP U1018 – Equipe Oncostat
SERA B.	INSERM SC10 - US19
TEGLAS J.P.	INSERM CESP U1018 - EQ05 - Epidémiologie clinique

Notes

Formulaire de pré-inscription

Vous pouvez également vous inscrire en ligne à l'adresse suivante :
https://eespe.vjf.inserm.fr
Nous attirons votre attention sur le fait que les demandes d'inscription en ligne seront traitées prioritairement aux inscriptions faites par courrier.

Nom .....
Prénom .....
Fonction .....
Date de naissance .....
Sexe ❑ M ❑ F ❑ Autre
Nom de la société : .....
Adresse professionnelle : .....
.....
CP .....Ville .....
Pays .....
.....
Tél prof. .....
Fax .....
E-mail .....
Tél portable .....
Adresse à laquelle le courrier doit être envoyé :
Nom .....
.....
CP ..... Ville .....
Pays .....

Photo

23 juin - 4 juillet 2025

Diplômes universitaires .....
.....
.....
.....
Situation professionnelle actuelle .....
.....
.....
.....
Avez-vous déjà suivi un enseignement en statistique ?
❑ oui ❑ non, si oui, le(s)quel(s) ? .....
.....
.....
Lieux ..... Année(s) .....
Avez-vous déjà suivi un enseignement en épidémiologie ?
❑ oui ❑ non, si oui, le(s)quel(s) .....
.....
.....
Lieu(x) ..... Année(s) .....

Avez-vous déjà suivi le STARC ? ❑ oui ❑ non,
Lieu(x) . . . . . Année(s) . . . . . .

Avez-vous déjà suivi un enseignement en santé publique ?
❑ oui ❑ non, si oui, le(s)quel(s) . . . . . .
Lieu(x) . . . . . Année(s) . . . . . .

Avez-vous déjà suivi un enseignement de l'école d'été ?
❑ oui ❑ non, si oui, le(s)quel(s) . . . . . .
Lieu(x) . . . . . Année(s) . . . .

Quel(s) logiciel(s) utilisez-vous pour analyser vos données ? . . . . . .
. . . . . .

Avez-vous déjà suivi un enseignement ou une formation en ligne ?
❑ oui ❑ non

Inscription :
- à titre individuel ❑
- en formation permanente Inserm ❑
- en formation permanente non Inserm ❑

Ce formulaire, accompagné d'une photo d'identité, et du justificatif de virement ou chèque du montant total des frais d'inscription, doit être retourné le plus rapidement possible avant le **6 juin 2025** à :

Ecole d'Eté de Santé Publique et d'Epidémiologie - INSERM CESP U1018
82 rue du Général Leclerc - 94276 Le Kremlin-Bicêtre Cedex (France)

Pour vérifier si les modules sont compatibles, vous pouvez aller sur le site internet de l'école d'été : https://eespe.vjf.inserm.fr

Choix du(des) module(s) :

❑ **1. Méthodologie de base en statistique et en épidémiologie**
F. Boufassa - 23 juin au 4 juillet, temps complet

❑ **2. Régression multiple en épidémiologie : régression linéaire, logistique et modèle de Cox**
J.-P. Empana et B. Heude - 23 au 27 juin, temps complet

❑ **3. Introduction à la modélisation des données longitudinales sous R : modèles marginaux et modèles mixtes**
H. Panjo et F. Artaud – 30 juin au 4 juillet, temps complet

❑ **4. Stratégie de recherche en épidémiologie**
M. Canonico - 23 au 25 juin, temps complet

❑ **5. Initiation à l'analyse de données avec R (tidyverse)**
R. Bauer et S. Guitton - 23 au 27 juin, après-midi

❑ **6. R avancé, automatiser ses rapports d'analyses avec R**
R. Bauer et S. Guitton – 30 juin au 4 juillet, après-midi

❑ **7. Introduction au *machine learning* en recherche biomédicale**
M. Sedki et J.-P. Teglas – 23 au 27 juin, temps complet