

STATISTIQUE ET ANALYSE DE DONNEES

Public visé par la formation et prérequis

Cette formation s'adresse à des professionnels de l'assurance ou de la finance qui souhaitent acquérir ou réviser les bases de la statistique mathématique (estimation, estimation par intervalle, tests) et de l'analyse de données (analyse en composantes principales, analyse factorielle des correspondances, classification), dans une optique d'application dans le monde de l'entreprise.

Prérequis : niveau mathématique équivalent à une licence 3 de mathématiques, connaissances élémentaires en probabilités et statistique.

Objectifs de la formation

A la fin de la formation, le stagiaire est capable de comprendre et mettre en œuvre des tests statistiques classiques (par exemple, test de significativité d'un coefficient dans un modèle linéaire, ou test d'adéquation à une distribution) dans des logiciels, ainsi que manuellement sur des exemples simples. Il est capable d'estimer les paramètres d'une loi (méthode des moments et maximum de vraisemblance), et de calculer des intervalles de confiance dans des cas classiques (modèle gaussien, proportions). Il maîtrise la mise en œuvre informatique sur le logiciel R de plusieurs méthodes d'analyse de données, l'interprétation des résultats, et peut avoir un regard critique sur l'adéquation de la méthode utilisée par rapport aux objectifs de l'analyse.

Contenus de la formation

La formation propose un cours théorique complet présentant les méthodes mathématiques utilisées (sans les preuves des théorèmes), des exercices et des cas d'application complets (travaux dirigés sur le logiciel R). Le stagiaire connaît ainsi le cadre théorique, et est capable d'aller chercher des approfondissements si nécessaires, ainsi que plusieurs exemples de mise en œuvre pratique sur différents jeux de données, dans des conditions proches de l'utilisation de la statistique dans un contexte professionnel.

1 - Statistique mathématique (5 séances)

- Notion de modèle statistique, modèles paramétriques et non paramétriques, introduction au modèle exponentiel
- Définition de l'estimateur et du risque quadratique (biais et variance de l'estimation), estimation par la méthode des moments, estimation par maximum de vraisemblance, estimation par intervalles
- Définition d'un test statistique (hypothèses, erreur de première espèce, puissance), hypothèse simple contre hypothèse simple, hypothèse simple contre hypothèses multiples
- Tests paramétriques : rapport des vraisemblances monotones, rapport des vraisemblances maximales. Mise en œuvre pour le modèle gaussien de tests sur l'espérance et sur la variance.
- Tests non paramétriques : test d'adéquation du χ^2 , test d'indépendance de deux distributions

2 - Analyse de données (5 séances)

- Rappels d'algèbre linéaire : espace vectoriel, espace affine, changements de repère, inertie.
- Analyse factorielle d'un nuage de points. Application à l'analyse en composantes principales (ACP).
- Prise en compte de variables discrètes : analyse factorielle des correspondances (AFC), analyse factorielle des correspondances multiples (ACM).
- Introduction à la classification. Méthodes de partitionnement (k-moyennes, nuées dynamiques).
- Classification ascendante hiérarchique (CAH).

Moyens et méthodes pédagogiques

Cours magistral complet au tableau pour l'exposé de la théorie mathématique, avec rappel de probabilités, de géométrie et d'algèbre linéaire. Exercices et présentation d'exemples d'application des méthodes sur logiciel. Les stagiaires reçoivent pour chaque partie du cours un polycopié rappelant les principaux résultats du cours, et des exemples complets illustrés, avec l'ensemble des jeux de données, procédures et commandes permettant de reproduire les applications sur leur propre ordinateur avec le logiciel R (libre et gratuit).

Évaluation de la formation / sanction de la formation

- Exercices à réaliser en dehors des sessions
- Evaluation sur les 10 questions écrites de l'épreuve d'accès au CEA
- Remise d'une attestation individuelle de formation

Formateur

Olivier Bonin est ingénieur des Ponts, des Eaux et des Forêts, docteur en statistique, et chercheur à l'IFSTTAR. Il est directeur adjoint du LVMT (laboratoire ville mobilité transport), et dirige un projet de recherche dans l'ITE Efficacy (institut de recherche public-privé). Il enseigne ou a enseigné la statistique et l'analyse de données en école d'ingénieur et à l'université, et mobilise des méthodes de statistique avancée dans ses propres travaux de recherche (modèles à variables latentes, structural equation models, confirmatory factor analysis, etc.).

Durée de la formation et modalités d'organisation

- 35 h réparties sur 10 demi-journées de mars à juin
- De 9 h à 12h30 ou de 14 h à 17h30
- Paris intra-muros
- ½ journée par semaine sur 10 semaines
- Une trentaine de stagiaires