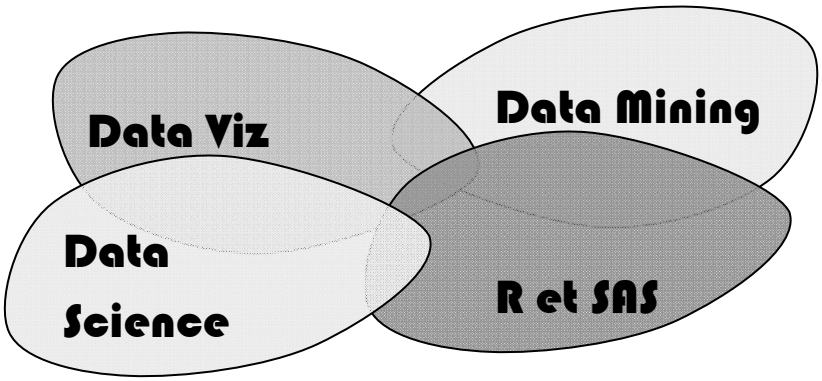
Olivier Decourt SARL



Catalogue formations 2018

Nous contacter par email :
formation@od-datamining.com

Visitez notre site : http://www.od-datamining.com/

Sommaire

Tarifs et conditions 5

R : maîtriser un logiciel souple, polyvalent et gratuit 6
[R_BASE] Initiation à R 7
[R_PLUS] Aller plus loin avec R 8
[R_REPORT] Construire des rapports avec R 9
[R_SHINY] Construire des applications interactives avec le package shiny 10

Initiations : premiers pas en statistique et en Data Mining 11
[STAT101] Initiation à la statistique 12
[DM] Qu'est-ce que le Data Mining ? 13
[IML] Initiation au langage SAS/IML 14

Formations au DataViz : présentation graphique des données 15
[VISUAL] Présenter clairement des données, construire des graphiques intelligents 16
[SAS_GRAPH1] Procédure SGPLOT 17
[SAS_GRAPH2] Modèles graphiques et GTL 18
[SAS_CARTO] Cartographie avec SAS 19
[R_GGPLOT] Produire des graphiques avec le package ggplot2 20

Formations statistiques : Data Science, Data Mining 21
[ANADON] Analyse des données 22
[GENMOD] Modèle linéaire généralisé 23
[MIXED] Analyse de la variance et modèles mixtes 24
[R_SCORING] Machine learning avec R 25
[REGQUALI] Régression sur variables qualitatives 26
[REGQUANTI] Régression sur variables quantitatives 27
[REPETE] Modélisation de données répétées 28
[SCORING] Panorama et comparaison des méthodes de scoring 29

Formations métier : biostatistique, marketing, actuariat 30
[BAYES] Introduction à la statistique bayésienne 31
[BIOSTAT] Modélisation biostatistique 32
[GLMTARIF] Modélisation pour la tarification en assurance 33
[POWER] Calcul de puissance et de nombre de sujets nécessaires 34
[SCOMKT] Scoring pour le ciblage marketing 35
[TYPOMKT] Typologie pour la segmentation client 36

Tarifs et conditions

Groupes de 1 à 8 personnes

Nous proposons un tarif unique en intra-entreprise, quel que soit le niveau du cours : 1 250 € HT par jour de formation.

Ce tarif est applicable à tout groupe de huit personnes maximum. Il inclut la réalisation et la distribution de supports de cours pour chaque participant.

Notre organisme de formation est déclaré auprès de la préfecture de région, et facture la TVA (20%).

Les cours peuvent se dérouler en anglais ou en français.

Formations sur mesure

Les contenus de cours qui sont détaillés dans les pages suivantes ne sont pas limitatifs. Il ne s'agit que des contenus **standards**.

Ceux-ci peuvent être **adaptés à vos besoins**, qu'il s'agisse de la modification d'un contenu existant ou de la création d'un nouveau cours. Ces modifications n'entraînent **aucune modification des tarifs ci-dessus**.

Les logiciels sur lesquels la formation peut s'appliquer sont indiqués à chaque cours. Si vous avez une demande sur un autre outil, n'hésitez pas à en faire la demande.

Cours en province

Les frais de déplacement et d'hébergement sur place du formateur seront facturés en sus sur présentation de justificatifs.

Supports

Les supports sont fournis à chaque participant sur des clés USB. Ces supports sont en français (ou, sur demande, en anglais) et mêlent théorie et applications.

R : maîtriser un logiciel souple, polyvalent et gratuit

[R_BASE] Initiation à R

Le logiciel R est principalement conçu pour des utilisations statistiques. Il recèle cependant de très nombreuses fonctionnalités de gestion de fichiers. Cette formation peut être couplée avec STAT101R, pour prolonger l'apprentissage avec la production de statistiques descriptives sous R.

Durée : 2 jours

Pré-requis : aucun

Logiciels possibles : R

1. Présentation de R

- Télécharger le logiciel
- Packages
- Environnement de base
- R Studio
- R Commander

2. Principes du langage R

- Fonctions : principes, utilisation, personnalisation
- Types de données
- Structures de données

3. Récupération de données

- Import d'un fichier texte
- Import d'une table SAS
- Import d'un fichier Excel

4. Vérification du contenu d'une table

- Statistiques descriptives
- Tableaux de fréquence
- Graphiques en bâtons
- Histogrammes et boxplots

5. Manipulation de données

- Création de variables, fonctions
- Empilement de tables
- Fusions de tables
- Transpositions et réorganisations des données

6. Sauvegardes

- Sauvegarde de données
- Export de données

[R_PLUS] Aller plus loin avec R

Ce stage permet de se perfectionner autour des fonctionnalités de packages de gestion de données (dplyr, reshape2), de la création de fonctions personnalisées et des data.tables.

Durée : 2 jours

Pré-requis : R_BASE

Logiciels possibles : R

1. Fonctions personnalisées et automatisation

- Construire une fonction
- Boucles
- Structures conditionnelles
- Utilisation de colnames pour l'automatisation
- Vectorisation (apply, sapply, tapply, lapply)

2. Gestion des données

- Package reshape2 : transpositions
- Package dplyr : requêtes
- Package forcats : gestion des facteurs
- Package lubridate : gestion des dates
- Gestion des doublons

3. Le package data.table

- Filtres
- Index
- Agrégats
- Jointures

[R_REPORT] Construire des rapports avec R

Ce stage permet d'explorer les différents packages de restitution automatique avec R. Il permet de produire des documents Word ou PowerPoint avec le package ReporteRs, des classeurs Excel avec xltabr, HTML, PDF et Word avec Rmarkdown.

Durée : 2 jours

Logiciels possibles : R 3.4 et +

Pré-requis : R_GGPLOT

1. Reporting avec Rmarkdown

- Principe de Rmarkdown
- En-tête
- Chunks
- Inline code
- Contenus : tableaux, graphiques

2. Reporting avec xltabr

- Rappels sur le package reshape2
- Export de tableaux croisés vers Excel

3. Reporting avec ReporteRs

- Principe de ReporteRs
- Insertion de textes
- Objet FlexTable
- Insertion de tableaux et de graphiques
- Mise en page
- Bookmarks : comment s'appuyer sur un document modèle

[R_SHINY] Construire des applications interactives avec le package shiny

Ce stage permet de s'initier au développement de mini-applications d'entreprise avec le package shiny : interfaces homme/machine, calculs et restitution.

Durée : 2 jours

Pré-requis : R_PLUS

Logiciels possibles : R

1. Principes de shiny

- Fonctionnement d'une application shiny
- Transmission d'informations
- Réactivité

2. Fonctions pour les calculs

- Scripts server.R et global.R
- Fonction shinyServer
- Fonction reactive
- Fonction validate
- Fonction isolate
- Fonction observe
- Mise à jour d'objets input

3. Fonctions pour l'interface

- Fonction de mise en page
- Exemples de mise en page
- Objets input
- Objets output
- Insérer des éléments

Initiations : premiers pas en statistique et en Data Mining

[STAT101] Initiation à la statistique

Ce stage est destiné aux personnes désireuses de découvrir les principes et les applications de la statistique. Il couvre principalement la statistique descriptive à une ou deux variables (graphiques et tableaux) et se termine sur un élargissement aux techniques plus avancées (tests, prévisions).

Ce cours est prévu sous forme appliquée ; seules les formules indispensables seront présentées.

Durée : 2 jours

Logiciels possibles : Excel, SAS, R, SAS Enterprise Guide

Pré-requis : manipulations de base du logiciel d'application

1. Vocabulaire et outils

- Distinguer qualitatif et quantitatif
- Modalité, cardinalité
- Population, champ d'une étude, individu

2. Résumer (une information)

- Notion de distribution
- Indicateurs numériques : moyenne, médiane, quantiles
- Dispersion : variance, écart-type, coefficient de variation
- Tableaux de fréquence
- Graphiques : bâtons, boxplots, histogrammes, courbes de densité

3. Comparer (deux informations ou plus)

- Graphiques : nuages de points, courbes
- Droite de tendance, lissage
- Graphiques : bâtons, radars
- Tableaux croisés
- Corrélations
- Indices

4. Extrapoler

- Sondages, échantillons, pondérations, redressement
- Tests statistiques : principe
- Test du khi-2, test de Student
- Modélisation (régressions)
- Traitement de séries chronologiques : désaisonnalisation, tendance

[DM] Qu'est-ce que le Data Mining ?

Une formation destinée aux chargés de projets et aux décideurs qui veulent savoir ce que recouvre exactement le mot de Data Mining. Quels sont les concepts, les démarches, les outils méthodologiques, les logiciels du marché avec leurs forces et leurs faiblesses ?

Durée : 1 jour

Pré-requis : aucun

Logiciels possibles : démonstrations sur R, SAS, Spad

1. Définition du Data Mining

- Un peu d'histoire
- Les domaines "historiques" d'application
- De nouveaux domaines d'expression

2. Les techniques du Data Mining

- La méthodologie
- Les arbres de décision
- Les régressions logistiques
- Les réseaux de neurones
- Les raisonnements à base de cas (MBR)
- Les machines à vecteurs-supports (SVM)
- Qu'est-ce qu'un score ?

3. L'offre logicielle

- Les pré-requis
- Les critères importants
- Quelques outils comparés

[IML] Initiation au langage SAS/IML

Le module SAS/IML donne accès à un langage spécifique pour la manipulation de matrices et les opérations qui leurs sont associées (diagonalisation, inversion, résolution de systèmes linéaires, etc.). A partir de SAS 9.3, ce module permet également des échanges avec le logiciel R.

Durée : 1 jour

Pré-requis : SAS BASE (formation Educasoft)

Logiciels possibles : SAS, SAS Enterprise Guide (code)

1. Présentation de SAS/IML

- Des matrices, des rappels d'algèbre
- Afficher une matrice

2. Charger une table SAS en matrice

- Utilisation d'une table SAS, chargement de la table
- Récupérer les noms des variables

3. Opérations sur les matrices

- Opérations matricielles et sur les éléments de la matrice
- Sous-ensemble d'une matrice
- Maximum, minimum

4. Fonctions et routines

- La fonction LOC
- Fonctions matricielles, fonctions mathématiques
- Fonctions de dénombrement, fonctions statistiques
- Conditions et boucles

5. Affichages

- Instruction MATTRIB
- Tris, calculs de statistisques

6. Exporter une matrice en table SAS

7. IML et automatisation

- Utilisation et création de macro-variables
- Modules
- Appel de code SAS et de code R

Formations au DataViz : présentation graphique des données

[VISUAL] Présenter clairement des données, construire des graphiques intelligents

Ce stage est destiné à tous ceux qui ont besoin de résumer des jeux de données par des graphiques, qu'il s'agisse de présentations, de tableaux de bord (*dashboards*) ou de recherche exploratoire.

Cette formation se concentre sur les différents types de graphiques et les moyens de leur donner une efficacité maximale. Elle pourra être couplée avec une formation aux graphiques SAS.

Cette formation peut être couplée avec STAT101 pour découvrir les indicateurs statistiques usuels et faire le lien avec leur diffusion efficace.

Durée : 1 ou 2 jours selon les attentes et le niveau du public

Pré-requis : aucun

Logiciels possibles : Excel, SAS, R

1. Principes de base pour une présentation efficace

- Principes du gestaltisme
- Attributs pré-attentifs

2. Présentation des tableaux

- Styles des cellules (nombres, pourcentages, etc.)
- Bordures, fonds de couleur : comment orienter la lecture d'un tableau
- Sources et titres

3. Différents types de graphiques pour représenter…

- … une variable qualitative
- … une variable quantitative
- … une série chronologique (évolution, tendance, lissage)
- … plusieurs variables simultanément

4. Efficacité d'un graphique

- Axes, légendes
- Choix des couleurs
- Treillis et graphiques multiples

[SAS_GRAPH1] Procédure SGPLOT

Ce stage est destiné aux chargés d'études souhaitant découvrir la nouvelle procédure SAS produisant des graphiques à partir de la version 9.2 : SGPLOT.

Cette formation présente également son complément, la procédure SGPANEL, et peut se poursuivre par une formation aux modèles écrits en GTL.

Durée : 1 ou 2 jours selon les attentes et le niveau du public

Pré-requis : SAS Base (formation Educasoft)

Logiciels possibles : SAS 9.2 et +

1. **ODS Graphics : un principe de fonctionnement et une instruction-clé**
   - Filière graphique historique vs filière « ODS Graphics » Java
   - Instruction ODS Graphics
   - Instructions ODS : choisir l'emplacement de l'image produite
2. **Graphiques sur données qualitatives**
   - Diagrammes en bâtons
   - « Dot plots »
   - Barres et courbes
3. **Graphiques sur données quantitatives**
   - Histogrammes, courbes de densité
   - Boxplots
   - Nuages de points, courbes, lissages
   - Droites de régression, intervalles de confiance
   - Zones colorées
4. **Titres, légendes, axes et textes**
   - Gérer des légendes
   - Insérer du texte fixe ou dynamique
   - Définir des axes
   - Utiliser des caractères spéciaux
5. **Plusieurs graphiques avec la procédure SGPANEL**
   - Gérer les variables catégorielles
   - SGPANEL et bloc BY : différences et complémentarité

[SAS_GRAPH2] Modèles graphiques et GTL

Ce stage est destiné aux chargés d'études qui voudraient produire des modèles réutilisables de graphiques ; pour cela, à partir de SAS 9.1, le GTL ou Graph Template Language, qui sous-tend le système ODS Graphics, permet de produire de belles sorties standardisées.

Cette formation indiquera selon les versions de SAS quelles syntaxes et quelles possibilités offre le GTL, un langage qui s'enrichit très rapidement au fil des versions.

Durée : 1 ou 2 jours selon les attentes et le niveau du public

Pré-requis : SAS_GRAPH1

Logiciels possibles : SAS 9.2 et +

1. Modèles de graphiques : le principe du GTL

- Procédures statistiques, modèles graphiques
- ODS PATH : où se trouvent les modèles ?
- Créer un modèle à partir de la procédure SGPLOT
- Utiliser un modèle avec la procédure SGRENDER

2. Paramétrer un modèle

- Paramètres dynamiques
- Macro-variables contenant du texte
- Macro-variables contenant des nombres

3. Principaux éléments d'un modèle

- Canevas (« lattice »)
- Elément graphique
- Axes

4. Principaux éléments graphiques

- Barres, boxplots, dot plots
- Nuages de points, courbes, lissages, régressions
- Zones colorées, quadrillage (« blockplot »)
- Légendes, titres
- Caractères spéciaux, mise en forme des éléments

5. Canevas dynamique

- Le couple LAYOUT DATALATTICE / LAYOUT PROTOTYPE
- Canevas dynamique vs macro-programme

[SAS_CARTO] Cartographie avec SAS

Ce stage est destiné aux chargés d'études (géomarketing, épidémiologie, sociologie, …) ayant à produire des cartes avec SAS.

Cette formation permet de se familiariser avec les cartes via la procédure SGPLOT -- et la procédure SGMAP à partir de SAS 9.4M5.

Durée : 1 journée

Logiciels possibles : SAS 9.3 et +

Nouveau contenu

Pré-requis : SAS_GRAPH1

1. Fonds de carte

- Fonds de carte fournis par SAS (bibliothèque MAPS et MAPSSAS)
- Fonds de carte GFK (bibliothèque MAPSGFK, SAS 9.4)
- Import de fonds de cartes externes (proc MAPIMPORT, étape Data)
- Création d'un identifiant géographique unique

2. Cartographie

- Données à représenter
- Fonction POLYGON
- Légende
- Bulles
- Textes
- Flèches
- Procédure SGMAP (SAS 9.4M5 et versions supérieures)

3. Manipulations sur les fonds de carte

- Fusion des éléments de la carte : proc GREMOVE
- Simplification d'un fond de carte : proc GREDUCE
- Plusieurs cartes côte à côte (zoom urbain, outremer) en GTL

[R_GGPLOT] Produire des graphiques avec le package ggplot2

Ce stage vise d'application des principes du dataviz avec le package ggplot2 : construction de graphiques pour l'exploration statistique et la publication.

Durée : 2 jours

Pré-requis : R_PLUS, VISUAL

Logiciels possibles : R

1. Le package ggplot2

- Principes
- Exporter un graphique, sauvegarde et recyclage

2. La fonction qplot

- Syntaxe de base
- Diagramme en bâtons
- Histogramme
- Boxplot
- Nuage de points
- Courbe

3. La fonction ggplot et ses compléments

- Grammaire des graphiques, les composants d'un graphique
- Données
- Esthétique
- Géométrie
- Traitement statistique
- Coordonnées
- Légendes
- Axes
- Thème

4. Eclatement

- Principe
- Eclatement

Formations statistiques : Data Science, Data Mining

[ANADON] Analyse des données

Ce stage est destiné aux chargés d'études qui désirent voir ou revoir les principes de l'analyse de données à la française (ACM, AFC, ACP) et surtout leur utilisation à travers SAS, R (package FactoMineR) ou Spad versions 6 à 8. On y aborde également la classification.

Durée : 2 jours

Pré-requis : STAT101

Logiciels possibles : SAS, SAS Enterprise Guide (code), R ou SPAD

1. L'analyse en composantes principales (ACP)

- Choix du nombre d'axes factoriels
- Nuages des individus et des variables
- Cercle des corrélations
- Rotation VARIMAX et ACP

2. L'analyse des correspondances multiples (ACM)

- Choix du nombre d'axes factoriels
- Nuages des individus et des variables
- Individus et variables supplémentaires

3. Typologies

- Classification ascendante hiérarchique
- Nuées dynamiques
- Méthode mixte de Wong
- Description des classes
- Modéliser l'appartenance aux classes pour réaffecter

[GENMOD] Modèle linéaire généralisé

Les modèles présentés ici font de la régression linéaire et de la régression logistique des cas particuliers. Les Modèles Linéaires Généralisés (MLG) se proposent d'étudier les variables dont la normalité est prise en défaut (coûts, fréquences d'évènements, ...) et proposent des outils puissants.

Durée : 2 jours

Pré-requis : STAT101

Logiciels possibles : SAS, SAS Enterprise Guide (code)

1. Principes de la régression

- Vocabulaire et concepts
- La régression linéaire
- La régression logistique
- Leurs points communs

2. Modèle linéaire généralisé

- Loi de Y
- Fonction de lien
- Qualité du modèle
- Analyse de la déviance
- Analyse des résidus et autres vérifications
- Syntaxe de la procédure GENMOD de SAS

3. Exemples de modèles linéaires généralisés

- Régression de Poisson
- Régression binomiale négative
- Régression Gamma

4. Données répétées et corrélées

- GLMM : des modèles très flexibles
- Choix d'une structure de corrélation
- Applications avec la procédure GLIMMIX de SAS

[MIXED] Analyse de la variance et modèles mixtes

L'étude des données avec une analyse de la variance se conduit d'ordinaire sur des facteurs considérés comme fixes : c'est à dire qu'on se limite dans l'analyse et l'inférence aux valeurs qui ont été collectées au cours de la constitution des données. Des facteurs aléatoires et un modèle mixte étendent de manière très importante la puissance des modèles d'analyse de variance, et facilitent également le traitement des données à mesures répétées

Durée : 2 jours

Pré-requis : STAT101

Logiciels possibles : SAS 9.2 et plus, SAS Enterprise Guide (code)

1. Analyse de variance, effets fixes et aléatoires

- Buts et hypothèses de l'analyse de variance
- Effet fixe et effet aléatoire
- Théorie et notations
- Panorama de l'offre SAS pour l'analyse de variance

2. Analyse de la variance à effets aléatoires

- Syntaxe de la procédure MIXED
- Détection graphique d'effets
- Quantification d'un effet aléatoire, calcul de moyennes ajustées
- Comparaison de groupes, ajustements pour les comparaisons multiples
- Intégration de variables fixes quantitatives
- Interactions : instructions LSMEANS et SLICE

3. Modèles mixtes généralisés

- Principe et théorie des modèles linéaires généralisés
- Syntaxe de la procédure GLIMMIX
- Régression logistique à effets aléatoires
- Régression de Poisson à effets aléatoires
- Régression Gamma à effets aléatoires

4. Analyse de variance sur données répétées

- Variabilité inter-sujets et intra-sujets
- Les principales structures de covariance
- Comparaison et choix de la structure la plus adaptée aux données

[R_SCORING] Machine learning avec R

Ce stage apprend à construire des scores et des modèles statistiques prédictifs avec R : statistique exploratoire supervisée, régression logistique, arbres de décision, comparaison de modèles.

Durée : 3 jours

Pré-requis : R_PLUS, REGQUALI

Logiciels possibles : R

1. Gestion des données

- Imputation
- Equilibrage
- Bases d'apprentissage, de validation et de test

2. Statistiques descriptives

- Graphiques
- Mesures de liaison
- Découpage en tranches

3. Modélisation

- Régression logistique
- Arbre de décision
- Analyse discriminante
- SVM
- Forêts aléatoires
- Réseaux de neurones (PMC)

4. Comparaison de modèles

- Indicateurs numériques
- Représentations graphiques
- Seuil optimal

[REGQUALI] Régression sur variables qualitatives

Destiné aux chargés d'étude s'intéressant à la modélisation d'une variable discrète (deux modalités ou davantage), ce stage permet de construire efficacement des modèles explicatifs et prédictifs (construction de scores).

Durée : 2 jours

Pré-requis : STAT101

Logiciels possibles : SAS, SAS Enterprise Guide (code), R, Spad, SPSS

1. Principe de la régression logistique

- Quelle est la forme des données à utiliser ?
- Lien avec la régression linéaire
- Les différentes fonctions de lien
- Mesurer la qualité d'un modèle logistique

2. La régression logistique à but descriptif

- L'analyse de la déviance, étude de l'impact d'une covariable
- Stratégies de construction de modèles cohérents
- Les coefficients
- Les odds-ratios
- La multicolinéarité

3. La régression logistique à but prédictif

- Qu'est-ce qu'un score ?
- La courbe ROC et le seuil optimal
- La courbe de lift
- Qualité d'ajustement

4. Etude d'une variable à plusieurs modalités

- Régression sur une variable ordonnée
- Régression sur une variable non ordonnée ou logit généralisé
- Application à la description d'une typologie

5. Modélisations alternatives d'une variable qualitative

- Analyse discriminante
- Réseaux de neurones
- Arbres de décision

[REGQUANTI] Régression sur variables quantitatives

Ce cours permet d'appréhender les principes de la régression, et sa mise en oeuvre. On y apprend le formalisme statistique associé, mais surtout la lecture des résultats, la détection d'erreurs et leur correction.

Durée : 2 jours

Pré-requis : STAT101

Logiciels possibles : SAS, SAS Enterprise Guide (code), R, SPSS

1. Découverte des données

- Distribution et normalité des variables
- Relations entre variables quantitatives
- Relations entre variables qualitatives

2. Régression linéaire simple

- Le modèle simple
- Sorties chiffrées
- Sorties graphiques

3. Sélection d'un modèle optimal

- Méthodes pas à pas
- Sélection sur un critère

4. Combattre la multicolinéarité

- Détecter la multicolinéarité
- Régression sur composantes factorielles
- Régression PLS
- Régressions contraintes : ridge et LASSO

5. Gestion des variables qualitatives

- Le modèle linéaire général, ANOVA, ANCOVA
- Choix de la modalité de référence

[REPETE] Modélisation de données répétées

Ce stage est destiné aux chargés d'études (biostatistique mais aussi actuariat ou marketing) qui ont à analyser des données composées de plusieurs mesures pour un même individu. Il peut s'agir de données répétées dans le temps (panels, visites au cours d'un essai clinique) ou non (mesures sur les 4 membres, sur les 2 yeux).

Cette formation utilise les procédures GENMOD, MIXED et GLIMMIX de SAS, en montrant comment les paramétrer et interpréter leurs sorties.

Durée : 1 ou 2 jours selon les attentes et le niveau du public

Pré-requis : REGQUANTI

Logiciels possibles : SAS, SAS Enterprise Guide (code)

1. Approche « modèle mixte » (GLMM)

- Effet aléatoire, corrélation, variance/covariance
- Structures de corrélation classiques
- Structures hétérogènes
- Structures « spatiales » pour intervalles irréguliers
- Choix d'une structure : critères d'Akaike et de Schwarz
- Influence de la structure sur les résultats de l'analyse

2. Approche « modèle linéaire à effets fixes » (GEE)

- GEE vs GLMM : ajustement marginal ou individuel ?
- Structures disponibles dans la procédure GENMOD
- Choix d'une structure : critère QIC
- Comparaison procédures MIXED / GLIMMIX et GENMOD

[SCORING] Panorama et comparaison des méthodes de scoring

Cette formation s'adresse aux chargés d'étude désirant avoir, en quelques jours, un aperçu technique et pratique des techniques usuelles de scoring. La formation s'achève avec une comparaison des forces et des faiblesses des différentes méthodes.

Durée : 3 jours

Pré-requis : DM

Logiciels possibles : SAS, SAS Enterprise Guide (code), SAS Enterprise Miner, R, Spad

1. Scoring avec les arbres de décision

- Principe général d'un arbre de décision
- Croissance et élagage
- Les principaux algorithmes : CHAID, CART, C4.5
- Arbres, bagging et boosting : comment rendre un arbre robuste
- Avantages et inconvénients

2. Scoring avec la régression logistique

- Modèle linéaire et modèle logistique
- Choix des variables, automatisation
- Coefficients et odds-ratios
- Courbe ROC, discrimination
- Avantages et inconvénients

3. Scoring avec l'analyse discriminante

- Approche géométrique
- Fonction linéaire discriminante
- Méthode DISQUAL : l'analyse discriminante sur données qualitatives
- Avantages et inconvénients

4. Scoring avec les réseaux de neurones

- Le neurone artificiel
- Apprentissage supervisé
- Lecture et interprétation des résultats
- Avantages et inconvénients

5. Autres méthodes de scoring

- Raisonnement basé sur la mémoire
- Machines à vecteurs supports (méthode Vapnik)
- Bagging et boosting
- Comparaison générale des méthodes de scoring

Formations métier :

biostatistique, marketing, actuariat

[BAYES] Introduction à la statistique bayésienne

Ce stage décrit les principes et les applications en épidémiologie des mécanismes de statistique bayésienne et ses différences avec la statistique classique (fréquentiste).

Des applications seront présentées en utilisant des logiciels libres (R et Winbugs).

Durée : 1 jour

Pré-requis : STAT101

Logiciels possibles : R

1. **Introduction à l'approche bayésienne**
2. **Lois a priori, vraisemblance, lois a posteriori, formule de Bayes**
3. **L'analyse bayésienne**
4. **Initiation aux méthodes MCMC (chaînes de Markov par Monte-Carlo), échantillonneur de Gibbs**
5. **Intervalles de crédibilité**
6. **Modèles de régressions bayésiens, DIC**
7. **Notions de tests bayésiens, facteur de Bayes**
8. **Comparaison des estimateurs bayésiens et fréquentistes**

[BIOSTAT] Modélisation biostatistique

Ce stage est destiné aux chargés d'étude œuvrant dans les essais cliniques, l'épidémiologie et la statistique animale. Il permet de faire le point sur la modélisation d'impact de facteurs sur une quantité d'intérêt à travers l'analyse de la variance et les modèles mixtes.

Cette formation peut être complétée par l'étude des données répétées pour l'intégration de plusieurs mesures sur un même individu dans le modèle.

Durée : 2 jours

Pré-requis : STAT101

Logiciels possibles : SAS, SAS Enterprise Guide (code)

1. Modèles possibles et hypothèses

- Analyse de variance / ANOVA
- Analyse de covariance / modèle linéaire général
- Modèle mixte
- Modèle mixte généralisé / GLMM

2. Analyse de la variance à facteurs fixes

- Normalité, une hypothèse indispensable
- Test de Fisher
- Tests de type 3
- Moyennes ajustées (LSMEANS)
- Comparaisons multiples et ajustement de la multiplicité

3. Analyse de la covariance à facteurs fixes

- Linéarité de la relation
- Interactions
- Moyennes ajustées en présence d'interactions : instructions LSMEANS et SLICE

4. Modèles mixtes

- Effet fixe vs effet aléatoire
- Estimation de l'effet aléatoire : part de variance, e-Blup
- Introduction à l'analyse de données répétées

[GLMTARIF] Modélisation pour la tarification en assurance

Ce stage est destiné aux chargés d'étude et actuaires qui ont à modéliser les sinistralités et coûts moyens en assurance non vie. Des applications sur les modèles composites (IARD) et de prime pure (santé) sont proposées sous SAS ou R.

Cette formation traite aussi de la phase exploratoire et de la modélisation de données répétées (GEE vs GLMM).

Durée : 2 jours

Pré-requis : STAT101

Logiciels possibles : SAS, SAS Enterprise Guide (code), R

1. Modèle Gamma pour le coût moyen

- Phase exploratoire : adéquation à la loi
- Phase exploratoire : lien log et variables quantitatives
- Construction et simplification du modèle
- Commentaire du modèle
- Cas particulier : modèle Gamma pour la prime pure

2. Modèle poissonnien pour la sinistralité

- Agrégation des données : comment et pourquoi ?
- Phase exploratoire : lien log et variables quantitatives
- Choix de la loi : Poisson ou binomiale négative ?
- Construction et simplification du modèle
- Commentaire du modèle
- Extension : modèles ZIP et ZINB

3. Modélisation des données répétées

- Approche GEE
- Approche GLMM
- Choix de la structure de corrélation
- Interprétation des résultats

[POWER] Calcul de puissance et de nombre de sujets nécessaires

Ce stage est destiné aux biostatisticiens désireux de calculer la puissance de leurs tests statistiques ou voulant quantifier le nombre sujets à inclure dans un essai clinique pour détecter un effet de manière significative.

Durée : 1 journée

Pré-requis : STAT101

Logiciels possibles : SAS, SAS Enterprise Guide (code)

1. **Rôle statistique du nombre de sujets dans les études cliniques**
2. **Méthodes d'estimation de la taille d'étude pour les tests simples**
3. **Application à des cas concrets, présentation de la proc POWER**
4. **Estimation dans des situations plus complexes : régression logistique, modèle de survie**
5. **Introduction du calcul de taille d'étude et de puissance *a posteriori* avec le modèle linéaire mixte**

[SCOMKT] Scoring pour le ciblage marketing

Ce stage est destiné aux chargés d'étude marketing qui ont besoin d'optimiser leurs ciblages pour des campagnes, de quantifier l'appétence à un produit, d'optimiser le cross-selling. Pour cela, ils utiliseront des techniques de modélisation classiques (régression logistique, arbres de décision) ou innovantes (modèle uplift).

Cette formation traite aussi de la phase exploratoire et de l'utilisation du modèle (courbe de lift, courbe ROC, suivi dans le temps).

Durée : 2 ou 3 jours selon les attentes et le niveau du public

Pré-requis : STAT101

Logiciels possibles : SAS, SAS Enterprise Guide (code), Spad, R

1. Définition de la problématique et préparation de la base

- Quelles problématiques se prêtent à un score ?
- Cerner la variable cible
- Pourquoi équilibrer un échantillon ?
- Scission en deux ou trois corpus de données

2. Statistique exploratoire

- Graphiques : courbes de densité, weight of evidence
- Indicateurs de liaison : Cramer, Fisher, Wilcoxon, Kolmogorov-Smirnov
- Caractérisation par les variables qualitatives

3. Mise au pas des données

- Ecrêtement de valeurs extrêmes
- Gestion de données manquantes
- Mise en tranches des variables quantitatives

4. Modélisations

- Régression logistique : sélection du meilleur modèle
- Régression logistique : simplification et interprétation du modèle
- Arbre de décision
- Régression logistique uplift

5. Comparaison et utilisation de modèles

- Stabilité du modèle : AUC sur les trois corpus de données
- Performance du modèle : courbe de lift

[TYPOMKT] Typologie pour la segmentation client

Ce stage est destiné aux chargés d'étude marketing qui ont besoin de bâtir des groupes homogènes de clients. Les techniques abordées sont la classification hiérarchique, les K-moyennes, ainsi que la méthode mixte ; une méthode proposée par Spad, la segmentation par arbre, est également présentée (dans ce logiciel, avec une macro SAS ou du code R).

Durée : 1 à 2 jours selon les attentes et le niveau du public

Pré-requis : STAT101

Logiciels possibles : SAS, SAS Enterprise Guide (code), Spad, R

1. Analyse factorielle

- Analyse exploratoire
- Création de facteurs synthétiques
- Influence du codage sur le résultat
- Sélection du nombre optimal de facteurs

2. Segmentation

- Classification Ascendante Hiérarchique, dendrogramme
- K-moyennes, CCC, formes fortes
- Méthode mixte / de Wong
- Classification explicite par arbres (sous Spad)

3. Description des classes, reclassement

- Description univariée (caractérisation)
- Réaffectation aux classes par arbre
- Réaffectation aux classes par régression logistique
- Réaffectation géométrique (distance aux centres)