

MULTIVARIATE ANALYSIS AND CLASSIFICATION (PCA, FCA, MCA, HCA, K-MEANS, DFA)

This course is intended for people who wish to master the concepts of multivariate factorial analysis and classification but also put in practice these methods.

The course is addressed to anyone seeking statistical knowledge on:

- The concepts of multivariate factorial analysis and classification
- The contexts in which each method is applied
- The mathematical (methodological) foundations of these analyses
- The implementation and interpretation of the results of the different methods

Duration: 28 hours (4 days)

Location: The course is available in-person at your premises or online via videoconference.

Price (excluding VAT) :

- Inter-company online training : n/a
- Intra-company online training : \$7650,00 per training
- Intra-company on site training : \$7650,00 per training + trainer's travel expenses

Trainee profiles:

- Technicians
- Employees
- Executives
- Researchers
- Students

Required experience:

Participants must have a good knowledge of basic statistical tools: correlation, standard deviation, variance, confidence intervals, hypothesis testing.

Training objectives

- Identify which multivariate analysis or classification methods to use depending on the context
- Mathematical concepts of these methods
- Implement multivariate factorial analyses: PCA, FCA, MFCA, DFA
- Interpret the results and graphs of multivariate analyses
- Evaluate the quality of the above statistical analyses
- Explain the difference between the concepts of contribution and cosine squared
- Explain the distances and aggregation methods best suited to the classification objectives
- Understand the differences between HCA and k-means
- Interpret the results of a classification method like HCA and k-means

Training syllabus

- **Introduction to the various methods of multivariate analysis**
 - Limitations of classical statistics
 - Fields of application of the different methods of multivariate analysis
 - Introduction to data mining - Description objectives - Prediction objectives
 - Dataset structure
 - Presentation of the range of methods:
 - Principal component analysis
 - Simple and multiple correspondence factorial analysis
 - Canonical correlation analysis
 - Discriminant factorial analysis
 - Classification methods: hierarchical ascending classification, k-means
 - General principles of the different methods - Notions of: Distance, Inertia and variance, Factorial axes
- **Concept of correlation**
 - Definition of the correlation coefficient
 - Interpretation of the value of the correlation coefficient
 - Sources of confusion: correlation, causation, slope ...
 - The different correlation coefficients: Pearson's coefficient - Spearman's coefficient
- **Implementation of a Principal Component Analysis (PCA)**
 - Dataset structure and application context
 - Detailed objectives of a PCA
 - Choice of the representation axes (choice of the number of principal components)
 - Interpretation of graphic outputs: factorial circle and graph of individuals
 - Interpretation of factorial axes
 - Contribution of individuals and variables to the axes
 - Quality of representation of individuals and variables on the axes: cosine squares
 - The different PCAs: standardized and non-standardized PCA - non-parametric PCA
 - Positioning of a categorical illustrative variable
 - Mix-ups and errors to be avoided
 - Quick presentation of Simple and Multiple Correspondence Analyses if desired
 - Differences between PCA and FCA
- **Implementation of a Factorial Correspondence Analysis (FCA)**
 - Dataset Structure: contingency table, individual data (qualitative variables)
 - Context of application and detailed objectives of FCA
 - Difference between PCA and FCA
 - FCA methodology: Chi² distance - Row profiles - Column profiles
 - Choosing representation axes (choosing the number of axes)
 - Interpretation of graphical outputs: modality graph

- Contributions and squared cosines of the modalities to the axes
- Errors to avoid
- **Implementation of a Multiple Correspondence Analysis (MCFA)**
 - Dataset structure
 - Context of application and detailed objectives of the MCFA
 - Difference between FCA and MCFA
 - MCFA methodology
 - Choosing the representation axes (choosing the number of axes)
 - Benzecri correction: corrected eigenvalues and explained inertia rates
 - Interpretation of the graphical outputs: Graph of the modalities - Graph of the variables
 - Contributions and squared cosines of the modalities to the axes
 - Errors to avoid
- **Implementation of a classification by a hierarchical ascending classification (HAC)**
 - Dataset structure
 - Context of application and detailed objectives of the HAC
 - Reading a dendrogram
 - Choosing the number of classes
 - Classification on individuals
 - Classification on variables
 - Classification on modalities
 - Choosing the distance according to the context: classification on individuals, variables or modalities
 - Choosing the linkage (jump, criterion of "rapprochement" between the classes): Ward's method, simple link, complete link ...
 - Interpretation of software outputs
- **Implementation of a k-means classification**
 - Presentation of the objectives of the k-means method
 - Advantages and disadvantages of HCA and k-means
 - Cluster determination
 - Presentation of the different versions of the algorithm
 - Using k-means as a complement to PCA
 - Classification on large data sets
 - Implementation tips
 - Interpretation of software outputs
- **Implementation of a Discriminant Factorial Analysis (DFA)**
 - Dataset structure and application context
 - Detailed objectives of DFA
 - Concepts of classification and discrimination
 - DFA Methodology
 - Comparison with PCA
 - Interpretation of software outputs: factorial circle, variable x-axis correlations
 - Quality of the DFA (of the discrimination obtained): Univariate and multivariate tests (Wilks' lambda), Graph of individuals, Confusion matrix (and possibly ROC curve)

- Errors to avoid

Training organization

Arkesys teaching team:

- **Thierry Anthouard** is the head of the Arkesys Group's statistical training program and has always been passionate about the field of statistics. In 1992, he launched the development of the Arkesys Group's statistics training program. His "by example" pedagogical approach allows him to popularize statistics and to make it accessible to all learners. As a consultant supporting key accounts, he adapts to all types of contexts and learning issues.
- **Jérôme-Philippe Garsi** is a statistical instructor with 13 years of experience in the training field. Since his doctorate on clinical issues, his work is mainly focused on the interest of populations, their health and well-being. At ease with any audience, he makes pedagogy and the simplification of scientific knowledge a priority. To do so, he always takes the greatest care to be clear in his written documents as well as in his oral presentations.

Teaching techniques:

- Trainees are welcomed in a dedicated virtual classroom
- Training materials are projected
- Supplemental exercises can be loaded onto a USB key

Follow-up: evaluating training participation and results

- Attendance sheets
- Review and evaluation exercises
- Theoretical explanations followed by guided practice and independent learning
- Evaluation at the end of the course
- Certificate of completion of the training course

Accessibility for disabled people:

People with disabilities who wish to take this course can contact us directly so we can examine together the best way to proceed.

Amount of time the course materials will be available before it starts: 2 weeks

Online training: The link to the virtual classroom will be sent by email the week before the course starts.

Contact: For further information you can contact us by email at: training@xlstat.com or by phone at (646) 412 3348.