# 1CC5000 – Statistics and Learning

**Instructors:** Arthur Tenenhaus
**Department:** DÉPARTEMENT MATHÉMATIQUES, MATHÉMATIQUES
**Language of instruction:** FRANCAIS, ANGLAIS
**Campus:** CAMPUS DE PARIS - SACLAY
**Workload (HEE):** 60
**On-site hours (HPE):** 36,00

**Description**

The objective of this course is to introduce the mathematical, methodological and computational bases of statistical inference from data. First, the principles and formalisms of mathematical statistics will be taught. This includes the definition of statistical models, the bases of estimation theory, the concepts of hypothesis testing.
Second, the basic methods and algorithms of statistical learning will be introduced, including supervised learning for regression and classification as well as unsupervised learning. Finally, the students will test several algorithms and libraries of statistical learning in practical classes with Python.

**Quarter number**
ST4

**Prerequisites (in terms of CS courses)**
Convergence-Integration-Probabilities

**Syllabus**
1. Random variables and samples, descriptive statistics, empirical measure.

2. Statistical models and problems of statistical inference

    a. Families of distributions and parametric models
    b. Exhaustive statistics, factorization theorem, exponential family.
    c. Regression Models

3. Parameter estimation.

    a. A few estimators: method of moments, maximum likelihood
    b. Properties of estimators (bias, consistency, risk, Cramer-Rao

bound, asymptotic properties, asymptotic normality, consistency and asymptotic normality of the ML estimator)
c. Central limit theorem, Delta method, Continuity theorem, Slutsky's theorem
d. Confidence regions

4. Bayesian Estimation : Bayes theorem, prior and posterior distributions, conjugate distributions, loss function and Bayesian point estimates.

5. Hypothesis Testing

a. General framework and method for testing statistical hypotheses : alternative hypotheses, risks and power, test statistics, rejection region, p-value
b. Parametric tests: Neyman-Pearson lemma, asymptotic tests.
c. Non-parametric tests (adequacy tests : chi2, Kolmogorov-Smirnov, Cramer Von-Mises ; population comparison tests: Wilcoxon)

6. Linear regression, generalized additive models, trees.

7. Model selection. L1-penalty (lasso) and L2-penalty (ridge regression), cross-validation.

8. Logistic model for classification.

9. An introduction to neural networks.

10. Principal Component Analysis. Unsupervised learning for clustering (K-means, hierarchical clustering)

**Class components (lecture, labs, etc.)**
11 x 1H30 lectures + 11 x 1H30 exercise classes + 2 x 3H practical classes + 2 x 1H30 Exams

**Grading**
2 compulsory tests, each accounting for 50% of the final mark.Test I: 1H30 wihout document, without electronic device, Mathematical Satistics, after 15H.Test II: 1H30 sans document ni matériel électronique, Statistical Learning, at the end of the course.

**Course support, bibliography**

- Lecture notes + exercise book
- Casella, G., & Berger, R. L. (2002). Statistical inference (Vol. 2). Pacific Grove, CA : Duxbury.

- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, pp. 241-249). New York: Springer series in statistics.

**Resources**

- Teaching staff (instructor(s) names): Arthur Tenenhaus, Laurent Le Brusquet, Julien Bect
- Maximum enrollment :
    - Exercise Classes :
        - 12 PCs with about 50 students (intermediate and advanced) in French or english
- Software : Practical classes in Python (with ScikitLearn, SciPy, StatsModels, Keras)
- Equipment-specific classrooms (specify the department and room capacity) : Normal classrooms with practical work on students laptops.

**Learning outcomes covered on the course**

At the end of the course, students will be able to :
- model a statistical inference problem
- estimate model parameters
- validate statistical hypotheses
- solve regression and classification problems from data
- identify homogeneous groups from data

# SEMESTER LONG COURSES