COURSE OF STUDY *Physics (LM-17)*
ACADEMIC YEAR *2024-2025*
ACADEMIC SUBJECT *Scientific Data Analysis Laboratory*

General information	
Year of the course	2nd
Academic calendar (starting and ending date)	1st semester: September – December 2024
Credits (CFU/ECTS):	6
SSD	FIS/01 – Experimental Physics
Language	English
Mode of attendance	Compulsory

Professor/ Lecturer	
Name and Surname	Alexis Pompili
E-mail	alexis.pompili@ba.infn.it , alexis.pompili@uniba.it
Telephone	+39 080 5442436
Department and address	Dipartimento Interateneo di Fisica (Office R15) - Via G.Amendola 173, 70125 Bari
Virtual room	Personal Zoom room
Office Hours (and modalities: e.g., by appointment, online, etc.)	Receiving hours for students: - Thursdays: 11-13 and 15-17 (@ office or online/zoom) - by appointment (@ office or online/zoom)

Work schedule			
Hours			
Total	Lectures	Hands-on (laboratory, workshops, working groups, seminars, field trips)	Out-of-class study hours/ Self-study hours
150	16	60	74
CFU/ECTS			
6	2	4	

Learning Objectives	Learners are expected to achieve, by the end of the course, a good knowledge about how to apply – critically – advanced statistics concepts in problems borrowed from real life research, by using proper statistical methods and techniques used in the field of Sub-nuclear and Nuclear Physics. Computational implementation of these methods and techniques is meant to be carried out within the established HEP toos/frameworks by using suitable code in C++ and/or python.
Course prerequisites	- Statistical Data Analysis Course (1st year Master degree course) - courses dealing with Programming languages during the Bachelor degree course (for instance Informatica).

Teaching strategies	Leveraging the theoretical statistics concepts learned in the Statistical Data Analysis Course, the students are guided, through a fully hands-on approach, to handle from basic to advanced methods and techniques of data analysis, with applications and examples borrowed from the High Energy Physics field. Students

	work, along the whole course, with their own personal account on a dedicated Linux virtual machine hosted by ReCas-Bari Data Center, properly configured and accessible also from home to allow extended individual exercise. In-class examples/exercises are fully supported by code files (in C/C^{++} and /or python) and complemented by slides-based presentations. All the needed files are provided; often some additional homework is proposed as further development of the exercise started in the laboratory lesson. A webpage is provided as additional support, with useful links to organize the given material and also some additional selected documentation, online manuals etc... Supplementary worksheet can be carried out individually but working in small teams is encouraged.
Expected learning outcomes in terms of	
Knowledge and understanding on:	● Understanding the scientific method, the nature, and the methods of research in Physics ● Knowledge of advanced computational techniques ● Knowledge of advanced mathematical tools commonly used in basic and applied research fields ● knowledge to configure an analysis of given (real or simulated) data and understanding the obtained results concerning the extraction or characterization of a physical signal or a physical trend, by means of a correct statistical treatment; ● knowledge of several statistical methods and techniques to handle data and understanding the context in which they can be properly applied and the approximations and uncertainties involved.
Applying knowledge and understanding on:	● Ability to identify the essential elements of a phenomenon ● Ability to use analogy to apply known solutions to new problems (problem solving) ● Ability to use analytical and numerical mathematical computation tools ● Ability to use electronic and computer technologies and their application to experimental data acquisition ● knowledge and ability to set up a statistically proper and computationally reliable data analysis task, by implementing suitable code (in C^{++} and/or python) within the established HEP tools/frameworks; ● Ability to build and test models and hypotheses.
Soft skills	● ***Making informed judgments and choices*** o Ability to work with increasing levels of autonomy, including taking responsibility in project planning and managing facilities. o Awareness of safety issues in laboratory activities o Develop the ability to autonomously design and set up a correct data analysis task and to evaluate the obtained results, also understanding how they can be eventually improved. ● ***Communicating knowledge and understanding***

	o Competence in communication in Italian and English in advanced fields of Physics o Develop computer skills related to experimental data processing/analysis and to how-to support their exposition/presentation, using specific and technical terminology; o communication and presentation skills - in English language - based on the specific terminology used in the field of experimental data analysis; moreover, dissemination of knowledge with appropriate scientific language; o general ability to work in a group, and to be inserted quickly and effectively in a workplace. ● ***Capacities to continue learning*** o Acquisition of basic knowledge tools for continuous learning and knowledge updates o Ability to learn and to transfer new techniques and methods in experimental data analyses; o Ability to handle and configure a data analysis - given a dataset of a wide variety in nature - by building a model and verifying/testing its validity and by extracting an underlying signature/feature and evaluating its significance in statistical terms.
Syllabus	

Content knowledge	The course contents and structure can be summarized as follow. The course aims to provide, at different levels of approach complexity, the knowledge and the related skills needed to configure, set up and handle, in a statistical and computational proper way, the analysis of a large variety of data, either real or simulated according to a certain model, and the evaluation of the results of this analysis, included the awareness of the implicit approximations and the comprehension of the statistical/systematic uncertainties and correlations involved. An overview of the suitable software tools and frameworks commonly established in the HEP sector are given, together with an extended use of them in a large set of examples of application and exercises. The course is designed into 3 modules (A, B, C) of increasing complexity. The hours can be divided among these 3 modules as follows: **A** : 14 (2 class + 12 lab) **B** : 46 (10 class + 36 lab) **C** : 16 (4 class + 12 lab) The usage of data analysis software tools is introduced gradually through the modules' sequence: **ROOT** in (A), **RooFit** package in (B), **Jupyter**, **Pandas** "ecosystem", and **RDataFrame** in (C). The detailed course program follows below. **Mod. A** Review of main useful **Unix/Linux** commands. Review of the main useful commands of the editors **vi** and **emacs**. Introduction to the **ROOT** data analysis software tool. Handling and representation of mathematical functions. Simple applications of representation of spreadsheet data using ROOT (TGraphErrors). Simple *random generation* of variables starting from a sampling distribution, storage into structures (histograms, *trees*), handling and visualization of generated data. Histograms' handling (operations with histograms), comparison bet-ween two or more histograms (*relative* normalization). Example of data-*Monte Carlo* (MC) comparison: between a data distribution and a set of different simulated components (to be composed into a *stacked* histogram); *absolute* normalization. Data-MC ratio; *rebinning* in presence of relevant fluctuations due to low statistics. Histogram of an efficiency and binomial bin error. Example of a simple binary classifier to extract a signal from a set of backgrounds: selection variables and evaluation of their background rejection power by comparing their ROC curves. Choice of the Working Point. **Mod. B**

	Introduction to the ROOT package **RooFit**. Maximum Likelihood interpolation method; implementation of a complex Probability Density Function and components of a theoretical fit model. *Extended* maximum likelihood. *Binned* and *unbinned* fits with examples. Introduction to the support of the modeling within RooFit; libraries, *workspace* and *factory* in RooFit. Background modelling: standard polynomials, *Chebyshev* and *Bernstein* polynomials, *Argus* function, polynomials functions with thresholds. Sigmoidal functions; example of fitting an efficiency curve with an *error* function. *Goodness-of-fit*: normalized chi-squared, fit probability and bin-by-bin pulls. Pulls as tool to check for biases and under/over-estimation of uncertainties. Examples of fits to invariant mass distributions; case of a particle /resonance with an intrinsic width much smaller than experimental mass resolution (one or more gaussians fit) and case of a particle/resonance with an intrinsic width of the same order of magnitude of the experimental resolution (*Voigtian* function, using a non-relativistic *Breit-Wigner* function). Fit with an explicit numerical convolution of a relativistic Breit-WIgner and gaussian resolution function. Empirical functions: *Crystal Ball* function (single- and double-sided) in case of radiative tails, *Johnson* function (as better alternative to more than two gaussians), *Landau* function. Exponential convoluted with a time resolution function (RooDecay model); inclusion of an event-by-event time error. Covariance/correlation matrix and estimation of the parameters' uncertainties with Hesse (parabolic approximation) and with Minos (possibly asymmetric errors) methods. Likelihood *scan* and *profile likelihood* with examples; verification of the equivalence between the method of likelihood *profiling* and Minos. Computation of the effective resolution in case of more than one gaussian in the fit, with estimation of its uncertainty by error propagation taking properly into account the correlations among fit parameters (getPropagatedError function). Simultaneous (multidimensional) fits in RooFit. Example: 2D fit of in-variant mass and proper time distributions (mass-lifetime fit of a particle). Non-parametric interpolation technique of the distribution of one (or more) random variable(s). Examples of application of the *kernel density estimation* (KDE) in one or two dimensions (RooKeysPdf and Roo2DKeysPdf). RooFit as a tool of distributions' generation according to chosen theoretical models (MC *toys*). Examples of generation of a distribution and its following fit; how to settle the *seed* of the random generator. *Closure test* of a fitting task. *Timing* evaluation of a *fitting task*. Kolmogorov-Smirnov test. Statistical significance of a signal by means of a likelihood ratio in the context of the Wilks theorem: example of approximate estimation. Brief mention to the statistical significance of a signal in the frequentist context (by using pseudo-experiments) resumed in the following module C.

	Review of the background subtraction methods: *sidebands* technique, *bin-wise method* and **sPlot** (by using the ROOT/TMVA package), with practical applications. Rejection of backgrounds and extraction of a signal. Example/exercise of a multivariate analysis by using a *boosted decision tree* (BDT)**.** Hypothesis testing and presence of a signal beyond background events. Computation of the *p-value*, and thus of the statistical significance, for the observation of a signal; definition of two models (null model for background-only hypothesis and alternative model for signal+background hypothesis) and, given the observed data, two calculation approaches: 1) *frequentist*, with a test statistics of *one-side Profile likelihood* type, by means of pseudo-experiments generation (to be executed in out-of-class hours, being computationally intensive), and 2) *asymptotic* (AsymptoticCalculator) with the same test statistics. Additional/optional exercise foresees the *p-value* computation as a function of a signal feature (for instance the signal mass), by using the first approach. **<u>Mod. C</u>** Introduction to **Python** scripting language: basic features, flux control, functions. Libraries for scientific computation: **numPy**, **sciPy, matplotlib, uproot (Pandas** "ecosystem"). Introduction to the **Jupyter** framework and examples of modern data analysis with signal extraction in the HEP field (extraction of a signal associated to a decaying particle by means of a full cut-based selection configured within the **Jupyter** framework).
Texts and readings	The material provided by the teacher covers the full course; it consists of a set of several pdf files inserted in a webpage dedicated to the course. This webpage contains also useful links on valuable online documentation and additional tutorials selected by the teacher and freely available on the internet. If needed by the topic, these files contain a brief and specific theoretical introduction/recall. As a further reference the student can consider the teaching material provided and the reference books suggested in the *Statistical Data Analysis* course, and in particular the well-known G. Cowan's (*) and L. Lista's (**) textbooks. Suggested reference for additionally exploring basic concepts of Machine Learning, at the end of module C, is the textbook by I. Goodfellow *et al*. (***). The examples and exercises carried out in the lab hours are run on a ReCas virtual machine dedicated to the course, where the students are given a personal account. The files to be used for the examples or to inspire (further) exercises are illustrated in the provided teaching material. Open data from the CMS experiment are used. (*) G. Cowan, Statistical Data Analysis, 1998, Clarendon Press. (**) L. Lista, *Statistical method for Data Analysis in Particle Physics*, 2019 (3r^{d} edition), Springer Verlag. (***) I. Goodfellow et al., Deep Learning, 2016, MIT Press.

Notes, additional materials	The proposed books are far richer than the conceptual content of the course. They can be used as a reference guideline concerning the basics concepts, methods and techniques used in the course.
Repository	**https://web2.ba.infn.it/~pompili/teaching.html**

Assessment	
Assessment methods	- Laboratory reports (20%), - Laboratory exam (80%)
Assessment criteria	The student is expected to have learned: - the conceptual and practical/technical knowledge of methods and techniques of data analysis, and the awareness of the proper statistics treatment and computational issues involved in the analysis; - the ability to design, configure and set up a data analysis task; - the knowledge and understanding of the statistically appropriate interpretation of the analysis results, including systema-tics uncertainties, eventual approximations, possible improvements.
Final exam and grading criteria	Evaluation/grading criteria for the final exam: ● *Knowledge and understanding: 3*0% ● *Applying knowledge and understanding:* 45% ● *Autonomy of judgment:* 10% ● *Communicating knowledge and understanding / Communication skills:* 10% ● *Capacities to continue learning: 5%*
Further information	
	The three modules are focused on laboratory applications, with examples and exercises, of statistical concepts introduced and studied in depth in the Statistical Data Analysis course. The in-class hours will be devoted to recall/refresh this knowledge but with concrete reference to the relevant functions, methods and algorithms that are involved and are available in the software tools/frameworks to be used. The very end of the third module foresees in-class hours to introduce a coherent and focused/guided overview of machine learning concepts and methods (slightly mentioned in the Statistical Data Analysis course) in the specific HEP context. It is implicitly assumed that a wider theoretical introduction and a wider range of applications in contexts other than HEP are provided in a dedicated Machine Learning course among those left as free students' choice.