UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

DIPARTIMENTO INTERUNIVERSITARIO DI FISICA

General informations about the course	
Course title	SCIENTIFIC DATA ANALYSIS LABORATORY
Degree course	Physics (Master degree)
ECTS	6
Compulsory Attendance	Yes
Teaching language	English

Professor/Lecturer	Alexis Pompili	alexis.pompili@ba.infn.it

ECTS details	Disciplinary Broad Area	SSD	ECTS
	Experimental Physics	FIS/01	6

Teaching activity & Time management by type	Period	Year	Lesson type
	1st semester	2nd year	Lessons (16h) Laboratory practise (60h)

Organizzazione della didattica	Total hours	In-lab/class hours	Out-of-class hours
	150	60/16	74

Course calendar	Starting date	Ending date
	Last week of september	Before Christmas break

Syllabus	
Prerequisits	Statistical Data Analysis Course (1st year Master degree course) Programming languages Course (Bachelor degree course)
Expected learning outcomes (According to Dublin Descriptors)	• **Knowledge and understanding** - knowledge to configure an analysis of given (real or simulated) data (not necessarily of HEP type) 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. • **Knowledge and understanding skills applied** - knowledge and ability to set up a statistically proper and computationally reliable data analysis task, by implementing suitable code (in C++ and python) within the established HEP toos/frameworks; - ability to build and test models and hypothesis. • **Autonomy of judgement** Ability to autonomously set up a correct data analysis task and to

	evaluate the obtained results, also understanding how they can be eventually improved. • **Transferable Communication skills** - computer skills related to experimental data processing/analysis and to how-to support their exposition/presentation, using specific and technical terminology; - 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; - general ability to work in a group, and to be inserted quickly and effectively in a workplace. • **Lifelong learning skills** - ability to learn and to transfer new techniques and methods in experimental data analysis; - 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 underlaying signature/feature and evaluating its significance in statistical terms.
Course contents summary and course structure	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 4 modules (A, B, C, D) of increasing complexity. The hours are divided among these 4 modules as follows: **A** : 12 (2 class + 12 lab) **B** : 40 (8 class + 36 lab) **C** : 18 (6 class + 12 lab) The usage of data analysis software tools are introduced gradually through the modules' sequence: **`ROOT`** in (A), **`RooFit`** and **`RooStats`** packages in (B), **`Jupyter`**, **`Pandas`** "ecosystem", and **`RDataFrame`** in (C).
Detailed Course Program	**Mod. A** Recall/Review of main useful **`Unix/Linux`** commands. Recall/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 between 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* histo-gram); *absolute* normlization. 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. **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 probabilità 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 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: `Hesse` (parabolic approximation) and `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 invariant mass and proper time distributions (mass-lifetime 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 1 or 2 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*. 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. Introduction to **`TMVA and RooStats`** packages of `ROOT/RooFit`. Review of the background subtraction methods: *sidebands* technique, *bin-wise method* and **`sPlot`** (by using **`TMVA`**), with practical applications. Rejection of backgrounds and extraction of a signal. Example/exercise of a multivariate analysis by using a *boosted decision tree* (BDT) within **`TMVA.`** 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 within `RooStats`: definition of 2 `ConfigModel` objects (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` class) 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 second approach.

	Mod. C 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).
Reference textbook and documentation material	**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 established 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, ClarendonPress. (**) L. Lista, *Statistical method for Data Analysis in Particle Physics*, 2019 (2nd edition), Springer Verlag. (***) I. Goodfellow et al., Deep Learning, 2016, MIT Press.
Note about the reference books	The books are far richer than the content of the course. They can be used as a reference guideline concerning the basics concepts, methods and techniques used in the course.
Teaching methods	Brief lectures and in-lab examples/exercises are fully supported by slide presentations and the needed files provided by the

	teacher. A webpage with useful links to additional selected documentation, online manuals etc… is provided as additional support. In-lab activities are meant to be rather interactive; students have to work individually but some sessions foresee working in small teams. 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 machine can be accessed also from outside the laboratory to allow students to use it in the out-of-class/study hours.
Evaluation method and assessment % of the final mark	- Laboratory reports (20%), - Laboratory exam (80%)
Evaluation criteria	The student is expected to have learned: - the conceptual and pratical 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 systematics uncertainties, eventual approximations, possible improvements.
Other notes	The first 3 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 end of 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 implicitely 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.