COURSE OF STUDY *Physics (LM-17)*

ACADEMIC YEAR *2023-2024*

ACADEMIC SUBJECT *Statistical Data Analysis*

General information	
Year of the course	1st
Academic calendar (starting and ending date)	1st semester: September – December 2023
Credits (CFU/ECTS):	6
SSD	FIS/01
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	40	15	95
CFU/ECTS			
6	5	1	

Learning Objectives	Learners are expected to achieve, by the end of the course, a good knowledge of advanced statistics concepts and methodologies widely used in the field of Sub-nuclear and Nuclear Physics. Moreover, they are expected to have acquired a critical approach to handle observations and measurements while being aware of the statistical and systematic uncertainties and the correlations involved.
Course prerequisites	Main/basic statistics' concepts learned/experimented in the Laboratory Courses of the Bachelor degree course.

Teaching strategies	Theoretical concepts are always complemented by practical applications and examples in order to establish a clear link between concepts on one hand and methodologies and application contexts on the other. Applications and examples are borrowed from the High Energy Physics field. Extended lectures and in-class examples/exercises are fully supported by slides presentations and the needed files provided by the teacher. A webpage, with useful links to additional selected documentation and online courses by the same

	authors of the reference books, is provided as additional support. Supplementary worksheets can be carried out individually but working in small teams is encouraged. Slides cover most of the course and provide to the students the hints about how to deepen their comprehension of a specific topic exploring the additional material and bibliographic references.
Expected learning outcomes in terms of	
Knowledge and understanding on:	o Understanding the scientific method, the nature, and the methods of research in Physics o Knowledge of advanced computational techniques o Knowledge of complex systems o Knowledge of statistical mechanics and statistical methods o Knowledge of advanced statistics concepts and methodologies useful to analyze experimental data and understand the context in which they can be properly used/applied and the possible approximations and uncertainties involved.
Applying knowledge and understanding on:	● Ability to use analogy to apply known solutions to new problems (problem solving) ● Ability to use analytical and numerical mathematical computation tools ● Ability to set up a statistically proper approach to a set of experimental data (not necessary deriving from HEP experiments); ● Ability to understand/propose the proper model for a physical process;
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 Develop the ability to autonomously design and set up a correct data analysis task, to evaluate the source and size of the statistical and systematic uncertainties that will be characterizing the resulting measurement, and to understand how they can be eventually reduced. ● ***Communicating knowledge and understanding*** o Competence in communication in Italian and English in advanced fields of Physics o communication and presentation skills, in English language, based on the specific terminology used in the field of statistics for physics and experimental data analysis; o dissemination of knowledge with appropriate scientific language. ● ***Capacities to continue learning*** o Acquisition of basic knowledge tools for continuous learning and knowledge updates

	o Deepen specific knowledge related to specific problems that could have to address in a master or a Ph.D. thesis.
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 of the concepts and methods to design and set up, in a statistical proper way, the analysis of a large variety of data, either real or simulated, according to a certain model, and the ability to evaluate the results of this analysis. The course is designed into 6 modules (A, B, C, D, E, F) of increasing complexity. The hours divided among these 6 modules and their reference titles are given as follows: **A: 6 – Theory of Probability** **B: 9 – Probability Density Functions of random variables** **C: 14 – Distribution functions & Central Limit Theorem** **D: 6 – Hypothesis testing** **E: 12 – Parameter estimation & Goodness-of-fit & Local statistical significance of a signal** **F: 8 – Classical confidence intervals & Global statistical significance of a new signal & Upper Limits** The detailed course program follows below. **<u>Mod. A</u>** Introduction to the Probability Theory. Probability and statistics. Probability and random variables. Different approaches to probability. Classical probability and example of the two dice roll. Probability as a relative frequency (frequentist interpretation). Subjective probability (Bayesian interpretation). Set theory and its representation applied to the sample space. Axiomatic Probability by Kolmogorov. Addition theorem: probability of the union of not disjoint event. Probability distributions and normalization condition. Joint and conditional probabilities. Independent events. Combination of detector efficiencies, detector in coincidence and test beam setup. Law of total probability (example of a scanning efficiency). Bayes' theorem and its extension through the law of total probability. Prior probability, likelihood and posterior probability. Examples from epidemiology: single and double screening. Example of particle identification and purity of a beam or a selected sample of particles. Frequentist use of the extended Bayes' theorem and recursive application.

	Mod. B Probability density functions; from discrete to continuous random variables; normalization condition. Cumulative distribution; quantile of order. Properties of PDFs: mode, median, expectation value, variance, skewness, kurtosis, (central) moments; standard deviation and standardized random variable. Vector of random variables. Joint PDF, marginal PDF, conditional PDF. Functions of random variables. Case of more than one random variable: covariance, correlation coefficient, covariance matrix. Independent and (un)correlated variables. Mixture of (sub-)samples: example of signal and background components. Propagation of variances: general rule and particular examples of sum, difference, product and ratio of two variables either correlated or uncorrelated. Physics example on how to build the covariance matrix in a lifetime measurement: decay time of a particle obtained by propagation of uncertainties. Orthogonal transformation of random variables with matrix notation; the case of two dimensions and the transformation of variables as rotation in the plane (diagonalization of the matrix). Physics examples: a) helix representation of the tracks in a uniform and axial magnetic field and transverse/longitudinal projections; distance/point of closest approach to the nominal production vertex; b) example of a 4-track system and background rejection in the Higgs to 4 leptons topology. **Mod. C** Bernoulli trial and *Binomial* distribution with applications; examples of the detection /reconstruction/selection efficiencies to be treated as binomial variables. *Multinomial* distribution with applications; example of the histogram for a fixed number of independent observations. *Poisson* distribution introduced for a stochastic process (radiative decay) and as a limit for the binomial distribution. Reproductive properties of binomial and Poisson distributions. Example of the histogram when the number of independent observations is itself a random variable. Example of forward-backward asymmetry with fixed or random number of measurements. Physics example of statistical and systematic uncertainties in the measurement of a branching fraction of a particle decay mode (Poissonian and binomial uncertainties); importance of the finite size of the Monte Carlo sample used to estimate the reconstruction efficiency (binomial distribution for efficiency estimate): the statistical error in Monte Carlo becomes a systematic error for the data measurement.

	Exponential distribution. *Uniform* distribution with examples (detection resolution with 1-strip clusters). Gaussian distribution as the limit of a Poisson distribution, standard gaussian PDF. Reproductive properties of the gaussian distribution. *Error* function and *Logistic* function with application (Covid-19 spread). N-dimensional generalization of the gaussian (*multivariate*) and *bivariate* normal distribution; error ellipses. *Log-normal* distribution. *Chi-square* distribution, number of degrees of freedom and Pearson theorem; physics example: the Maxwell-Boltzmann distribution for a gas. *Cauchy* (*Breit-Wigner*) distribution; truncation technique for pathologic functions. Convolution of a Breit-Wigner with a (gaussian) Resolution function and discussion of physics cases (intrinsic width versus experimental mass resolution). Convolution of an exponential with a resolution function (proper time exponential distribution convoluted with gaussian function centered at time zero). Landau distribution for the energy loss of a charged particle crossing a layer of matter of a given thickness. Markov and Chebyshev inequalities. Convergence criteria, weak and strong laws of large numbers. Definition and properties of the characteristic function and applications. Central Limit Theorem (CLT) and its main corollary; its demonstration using the characteristic function of a gaussian. Discussion of the validity of the CLT and of the requisites for his application. Physics example: multiple scattering and deviations at large angles (back-scattering). **Mod. D** Introduction to statistical tests. Null and alternative hypothesis in the simple case of a binary classifier. Critical and acceptance regions. Significance level of the test and decision boundary. Errors of the first kind and of the second kind (contamination), power of the test. Physics examples: separation in two classes, particle selections; efficiency and purity. Neyman-Pearson lemma, likelihood ratio. Construction of a test statistic and introduction to artificial neural networks and Machine Learning. ROC curve and application of the ROC curves as comparison among algorithms with physics examples. Understanding the Area-Under-Curve (AUC) and other figures of merit. Choice of the Working Point. **Mod. E** Introduction to parameter estimation, the process of inference. Parameters of interest and nuisance parameters. Measurements and their uncertainties: central value and uncertainty interval. Statistical and systematic uncertainties: precision and accuracy. Frequentist inference:

	confidence level and coverage. Estimators and PDF of an estimator. Example of a simple estimator in a gaussian case. Properties of estimators: consistency, unbiasedness and robustness. Minimum variance (Cramer-Rao) bound and efficiency of an estimator. Simple cases of a sample of N measurements: estimator for the expectation value (sample mean), estimator for the variance (modified sample variance), for the covariance and for the correlation coefficient. Introduction to the methods of building an estimator. Maximum Likelihood (ML) method; *Likelihood* and *Extended Likelihood* functions with examples (with unbinned and binned data samples). Gaussian likelihood functions, bias of the ML estimate of a gaussian variance. Variance of a ML estimator (RCF bound and graphical method). Errors with the ML method; second derivatives matrix (Hesse), likelihood scan and asymmetric errors (Minos). Properties of ML estimators. Minimum chi-square and Least Squares method (LS) and the easiest case of Linear Regression. Application of minimum chi-square for (binned) histograms in the gaussian approximation and when a Poissonian model has to be applied (when number of entries per bin is small); modified LS method. Introduction to the problem of combining two (or more) measurements; simultaneous fits and control regions; weighted average by application of the minimum chi-square in the gaussian approximation and assuming uncorrelation. Extension to the correlated case. Goodness-of-fit tests and p-value as observed significance level. Statistical significance of an observed signal (case of Poisson variables); Pearson chi-square test. Alternative approach with a likelihood ratio as a test statistic in the search for a new signal; example of a simple event counting experiment. Neyman-Pearson lemma. Wilks theorem and its applicability; example of likelihood-ratio by considering the background-only and signal-plus-background hypotheses. Validity of the Cowan asymptotic formulae and its verification with the pseudo-experiment (toys) method. Local statistical significance level and observation (discovery) or evidence of a signal. **Mod. F** Introduction to the classical confidence intervals (CI). Neyman CI: construction of the confidence belt and its inversion. Confidence level as probability coverage. CI for a gaussian distributed estimator. CI for the mean of the Poisson distribution. Confidence intervals using the likelihood function or the chi-square in the large sample limit. Binomial intervals and application of the Clopper-Pearson method.

	Rare signals and introduction to frequentist Upper Limits (ULs); example of a counting experiment. Unified Feldman-Cousins approach: how to avoid the flip-flopping issue and ensure the correct coverage; transition from central intervals to ULs. The issue of the dependence of upper limits on the expected amount of background in case of zero signal events. Modified frequentist approach for ULs: the CLs method with examples. How to read a Brazil Plot. Upper limits using the Profile Likelihood. Global statistical significance and the Look-Elsewhere-Effect (LEE) dilution. Method of Trial Factors; validity of the Gross-Vitells approximation and its verification with a scanning technique based on Monte Carlo toys; example of simplified calculation of a global p-value.
Texts and readings	For what concerns the Reference textbooks and the documentation material, the material provided by the teacher covers the full course; it consists of a set of several pdf files (in slides format) inserted in a webpage dedicated to the course. This webpage contains also useful links on valuable online documentation selected by the teacher and freely available on the internet. As further reference the student can consider the reference books suggested in the following list and used as a guide in the design and set up of the present course: 1. G. Cowan, *Statistical Data Analysis*, 1998, Clarendon Press; available in the library and also paper book provided by the teacher upon request. 2. L. Lista, *Statistical methods for Data Analysis in Particle Physics*, 2020 (2nd or 3rd edition), Springer Verlag, available in the library and also paper book provided by the teacher upon request. 3. W. J. Metzger, *Statistical Methods in Data Analysis*, 2010 (2nd edition), for the Radboud University of Nijmegen; free online. 4. F. James, *Statistical methods in Experimental Physics*, 2006 (2nd edition), World Scientific; paper book provided by the teacher upon request.
Notes, additional materials	The books are richer than the content of the course. They can be used as a reference guideline concerning the concepts and methods introduced and discussed in the course. Sometimes they allow further in-depth analysis useful for the supplementary worksheet. They offer a wide variety of examples and applications. However, a relevant part of the examples and applications presented in this course are original re-working of existing material and direct research experience in

	the HEP field and can be found only in the material provided by the teacher.
Repository	**https://web2.ba.infn.it/~pompili/teaching.html**

Assessment	
Assessment methods	The choice between the two following options is left to any student who follows the course: <u>Option A</u> - Oral exam (100%) … or <u>Option B</u> - Oral exam (50%), - Supplementary worksheet with a final seminar (50%) [can be prepared individually or in team (in the latter case individual contributions must clearly appear)]
Assessment criteria	The student is expected to have learned: - the conceptual knowledge and understanding of the advanced statistics concepts and methodologies useful to analyze experimental data; - the awareness of the features and possible issues of a proper statistics treatment involved in big data analyses; - the ability to design and set up the analysis of experimental data; - the knowledge and understanding of the statistically appropriate interpretation of the analysis results, including statistical and systematics uncertainties, the involved approximations, the possible improvements.
Final exam and grading criteria	Final exam is carried out as in options A or B according to student's choice. **Evaluation/grading criteria:** ● *Knowledge and understanding:* 40% (A) – 30% (B) ● *Applying knowledge and understanding:* 30% (A) – 20%(B) ● *Autonomy of judgment:* 10% (A) – 10% (B) ● *Communicating knowledge and understanding / Communication skills:* 10%(A) – 20% (B) ● *Capacities to continue learning:* 10%(A) – 20% (B)
Further information	
	Machine and Deep Learning concepts and methods are briefly introduced in a sort of overview for completeness purposes. However, it is implicitly assumed that dedicated courses will be foreseen for the interested students in the master course or, later, as PhD course. Nonetheless this course provides all the knowledge basis needed for such kind of dedicated courses.