CORSO DI STUDIO *Physics (LM-17)*

ANNO ACCADEMICO *2024-2025*

DENOMINAZIONE DELL'INSEGNAMENTO *Statistical Data Analysis*

Principali informazioni sull'insegnamento	
Anno di corso	1°
Periodo di erogazione	1° semestre: Settembre - Dicembre 2024
Crediti formativi universitari (CFU/ECTS):	6
SSD	FIS/01
Lingua di erogazione	Inglese
Modalità di frequenza	Obbligatoria

Docente	
Nome e cognome	Alexis Pompili
Indirizzo mail	alexis.pompili@ba.infn.it , alexis.pompili@uniba.it
Telefono	+39 080 5442436
Sede	Dipartimento Interateneo di Fisica (Ufficio R15) - Via G.Amendola 173, 70125 Bari
Sede virtuale	Zoom Room personale
Ricevimento	Ore di ricevimento: - Gioved': 11-13 e 15-17 (Ufficio R15 oppure online/su zoom) - per appuntamento (Ufficio R15 oppure online/su zoom)

Organizzazione della didattica			
Ore			
Totali	Didattica frontale	Pratica (laboratorio, campo, esercitazione, altro)	Studio individuale
150	40	15	95
CFU/ECTS			
6	5	1	

Obiettivi formativi	Ci si aspetta che gli studenti raggiungano, alla fine del corso, una buona conoscenza dei concetti e delle metodologie statistiche avanzate ampiamente utilizzate nel campo della fisica sub-nucleare e nucleare. Inoltre, ci si aspetta che abbiano acquisito un approccio critico per gestire le osservazioni e le misure, essendo consapevoli delle incertezze statistiche e sistematiche e delle correlazioni coinvolte.
Prerequisiti	Concetti principali/di base della statistica appresi/sperimentati nei corsi di laboratorio del corso di laurea triennale.

Metodi didattici	I concetti teorici sono sempre integrati da applicazioni pratiche ed esempi, al fine di stabilire un chiaro collegamento tra i concetti da un lato e le metodologie e i contesti applicativi dall'altro. Le applicazioni e gli esempi sono presi in prestito dal campo della fisica delle alte energie. Le lezioni estese e gli esempi/esercizi in classe sono completamente supportati da presentazioni di diapositive e dai file necessari forniti dal docente. Come ulteriore supporto viene fornita una pagina web con link utili a documentazione aggiuntiva selezionata e a corsi online degli stessi autori dei libri di riferimento. I fogli di lavoro supplementari possono essere svolti individualmente, ma si incoraggia il lavoro in piccoli gruppi.

	Le diapositive coprono la maggior parte del corso e forniscono agli studenti suggerimenti su come approfondire la comprensione di un argomento specifico esplorando il materiale aggiuntivo e i riferimenti bibliografici.

Risultati di apprendimento previsti ***Da indicare per ciascun Descrittore di Dublino (DD=*** ***DD1* Conoscenza e capacità di comprensione** ***DD2* Conoscenza e capacità di comprensione applicate** ***DD3-5* Competenze trasversali**	***- DD1: conoscenza e capacità di comprensione*** o Comprensione del metodo scientifico, della natura e delle modalità della ricerca in Fisica o Conoscenza delle tecniche di calcolo avanzate o Conoscenza dei sistemi complessi o Conoscenza della meccanica statistica e dei metodi statistici o Conoscenza dei concetti e delle metodologie statistiche avanzate utili per analizzare i dati sperimentali e comprendere il contesto in cui possono essere correttamente utilizzati/applicati e le possibili approssimazioni e incertezze coinvolte. ***- DD2: capacità di applicare conoscenza e comprensione*** ● Capacità di utilizzare lo strumento dell'analogia per applicare soluzioni conosciute a problemi nuovi (problem solving) ● Capacità di utilizzo di strumenti di calcolo matematico analitico e numerico ● impostare un approccio statisticamente corretto a un insieme di dati sperimentali (non necessariamente derivanti da esperimenti HEP); ● capacità di comprendere/proporre il modello corretto per un processo fisico; **- DD3-5: competenze trasversali** ***-DD3** :* ● **Autonomia di giudizio** o Capacità di lavorare con crescenti gradi di autonomia, anche assumendo responsabilità nella programmazione di progetti e nella gestione di strutture o sviluppare la capacità di progettare e impostare autonomamente un corretto compito di analisi dei dati, di valutare la fonte e l'entità delle incertezze statistiche e sistematiche che caratterizzeranno la misura risultante e di capire come possono essere eventualmente ridotte. ***-DD4**:* ● **Abilità comunicative** o Competenze nella comunicazione in lingua italiana e in lingua inglese nei settori avanzati della Fisica o capacità di comunicazione e presentazione, in lingua inglese, sulla base della terminologia specifica utilizzata nel campo della statistica per la fisica e dell'analisi dei dati sperimentali; o divulgazione delle conoscenze con un linguaggio scientifico appropriato. *- **DD5**:* ● **Capacità di apprendere in modo autonomo**

	o Acquisizione di strumenti conoscitivi di base per l'aggiornamento continuo delle conoscenze o consentire allo studente di approfondire conoscenze specifiche relative a problemi particolari che potrebbero essere affrontati in una tesi di master o di dottorato.
Contenuti di insegnamento (Programma)	I contenuti e la struttura del corso possono essere riassunti come segue. Il corso si propone di fornire, a diversi livelli di complessità dell'approccio, la conoscenza dei concetti e dei metodi per progettare e impostare, in modo statisticamente corretto, l'analisi di una grande varietà di dati, reali o simulati, secondo un determinato modello, e la capacità di valutare i risultati di questa analisi. Il corso è strutturato in 6 moduli (A, B, C, D, E, F) di complessità crescente. Le ore suddivise tra questi 6 moduli e i loro titoli di riferimento sono riportati di seguito: A: 6 - Teoria della probabilità B: 9 - Funzioni di densità di probabilità di variabili casuali C: 14 - Funzioni di distribuzione e Teorema del limite centrale D: 6 - Test di ipotesi E: 12 - Stima dei parametri & *Goodness-of-fit* & Significatività statistica locale di un segnale F: 8 - Intervalli di confidenza classici & Significatività statistica globale di un nuovo segnale & Limiti superiori Il programma dettagliato del corso è riportato di seguito. <u>Mod. A</u> Introduzione alla teoria della probabilità. Probabilità e statistica. Probabilità e variabili casuali. Diversi approcci alla probabilità. Probabilità classica ed esempio del lancio di due dadi. Probabilità come frequenza relativa (interpretazione frequentista). Probabilità soggettiva (interpretazione bayesiana). Teoria degli insiemi e sua rappresentazione applicata allo spazio campionario. Probabilità assiomatica di Kolmogorov. Teorema dell'addizione: probabilità dell'unione di eventi non disgiunti. Distribuzioni di probabilità e condizione di normalizzazione. Probabilità congiunte e condizionali. Eventi indipendenti. Combinazione di efficienza del rivelatore, rivelatore in coincidenza e configurazione del fascio di prova. Legge della probabilità totale (esempio di efficienza di scansione). Teorema di Bayes e sua estensione attraverso la legge della probabilità totale. Probabilità prioritaria, verosimiglianza e probabilità posteriore. Esempi di epidemiologia: screening singolo e doppio. Esempio di identificazione di particelle e purezza di un fascio o di un campione

	selezionato di particelle. Uso frequentista del teorema di Bayes esteso e applicazione ricorsiva. Mod. B Funzioni di densità di probabilità; da variabili casuali discrete a continue; condizione di normalizzazione. Distribuzione cumulativa; quantili di ordine. Proprietà delle PDF: modalità, mediana, valore dell'aspettativa, varianza, sbandamento, curtosi, momenti (centrali); deviazione standard e variabile casuale standardizzata. Vettori di variabili casuali. PDF congiunta, PDF marginale, PDF condizionale. Funzioni di variabili casuali. Caso di più variabili casuali: covarianza, coefficiente di correlazione, matrice di covarianza. Variabili indipendenti e (non) correlate. Miscela di (sotto)campioni: esempio di componenti di segnale e di fondo. Propagazione delle varianze: regola generale ed esempi particolari di somma, differenza, prodotto e rapporto di due variabili correlate o non correlate. Esempio di fisica su come costruire la matrice di covarianza in una misura di durata: tempo di decadimento di una particella ottenuto per propagazione delle incertezze. Trasformazione ortogonale di variabili casuali con notazione matriciale; il caso di due dimensioni e la trasformazione delle variabili come rotazione nel piano (diagonalizzazione della matrice). Esempi di fisica: a) rappresentazione elicoidale delle tracce in un campo magnetico uniforme e assiale e proiezioni trasversali/longitudinali; distanza/punto di massimo avvicinamento al vertice di produzione nominale; b) esempio di un sistema a 4 tracce e reiezione del fondo nella topologia del decadimento del bosone di Higgs in 4 leptoni. Mod. C Processo di Bernoulli e distribuzione binomiale con applicazioni; esempi di efficienza di rilevamento/ricostruzione/selezione da trattare come variabili binomiali. Distribuzione multinomiale con applicazioni; esempio di istogramma per un numero fisso di osservazioni indipendenti. Introduzione della distribuzione di Poisson per un processo stocastico (decadimento radiativo) e come limite della distribuzione binomiale. Proprietà riproduttive delle distribuzioni binomiale e di Poisson. Esempio di istogramma quando il numero di osservazioni indipendenti è esso stesso una variabile casuale. Esempio di asimmetria avanti-indietro con numero fisso o casuale di misure. Esempio di fisica: incertezze statistiche e sistematiche nella misura della frazione di ramificazione di una modalità di decadimento di una particella (incertezze poissoniane e binomiali); importanza della dimensione finita del campione Monte Carlo utilizzato per stimare l'efficienza di ricostruzione (distribuzione binomiale per la

	stima dell'efficienza): l'errore statistico di Monte Carlo diventa un errore sistematico per la misura dei dati. Distribuzione esponenziale. Distribuzione uniforme con esempi (risoluzione di rivelazione di una particella con cluster ad 1 microstrip in rivelatore a silicio o segmentati). Distribuzione gaussiana come limite di una distribuzione di Poisson, PDF gaussiana standard. Proprietà riproduttive della distribuzione gaussiana. Funzione di errore e funzione logistica con applicazione (diffusione di Covid-19). Generalizzazione n-dimensionale della distribuzione normale gaussiana (multivariata) e bivariata; ellisse degli errori. Distribuzione log-normale. Distribuzione Chi-quadro, numero di gradi di libertà e teorema di Pearson; esempio di fisica: la distribuzione di Maxwell-Boltzmann per un gas. Distribuzione di Cauchy (Breit-Wigner); tecnica di troncamento per le funzioni patologiche. Convoluzione di una Breit-Wigner con una funzione di risoluzione (gaussiana) e discussione di casi fisici (ampiezza intrinseca contro risoluzione sperimentale della massa). Convoluzione di un esponenziale con una funzione di risoluzione (distribuzione esponenziale a tempo proprio convoluta con una funzione gaussiana centrata al tempo zero). Distribuzione di Landau per la perdita di energia di una particella carica che attraversa uno strato di materia di un dato spessore. Disuguaglianze di Markov e Chebyshev. Criteri di convergenza, leggi dei grandi numeri deboli e forti. Definizione e proprietà della funzione caratteristica e applicazioni. Teorema del limite centrale (CLT) e suo principale corollario; dimostrazione utilizzando la funzione caratteristica di una gaussiana. Discussione della validità del CLT e dei requisiti per la sua applicazione. Esempio di fisica: diffusione multipla e deviazioni a grandi angoli (*back-scattering*). <u>Mod. D</u> Introduzione ai test statistici. Ipotesi nulla e alternativa nel caso semplice di un classificatore binario. Regioni critiche e di accettazione. Livello di significatività del test e confine decisionale. Errori del primo tipo e del secondo tipo (contaminazione), potenza del test. Esempi di fisica: separazione in due classi, selezione di particelle; efficienza e purezza. Lemma di Neyman-Pearson, rapporto di verosimiglianza. Costruzione di un test statistico e introduzione alle reti neurali artificiali e al Machine Learning. Curva ROC e applicazione delle curve ROC come confronto tra algoritmi con esempi di fisica. Comprensione dell'Area-Under-Curve (AUC) e di altre figure di merito. Scelta del punto di lavoro. <u>Mod. E</u> Introduzione alla stima dei parametri, il processo di inferenza. Parametri di interesse e parametri di nuisance. Misure e loro incertezze: valore

	centrale e intervallo di incertezza. Incertezze statistiche e sistematiche: precisione e accuratezza. Inferenza frequentista: livello di confidenza e copertura. Stimatori e PDF di uno stimatore. Esempio di uno stimatore semplice nel caso gaussiano. Proprietà degli stimatori: consistenza, imparzialità e robustezza. Vincolo di varianza minima (Cramer-Rao) ed efficienza di uno stimatore. Casi semplici di un campione di N misure: stimatore per il valore di aspettativa (media campionaria), stimatore per la varianza (varianza campionaria modificata), per la covarianza e per il coefficiente di correlazione. Introduzione ai metodi di costruzione di uno stimatore. Metodo della massima verosimiglianza (ML); funzioni di verosimiglianza e di verosimiglianza estesa con esempi (con campioni di dati non filtrati e filtrati). Funzioni di verosimiglianza gaussiane, bias della stima ML di una varianza gaussiana. Varianza di uno stimatore ML (limite RCF e metodo grafico). Errori con il metodo ML; matrice delle derivate seconde (Hesse), scansione della verosimiglianza ed errori asimmetrici (Minos). Proprietà degli stimatori ML. Minimo chi-quadro e metodo dei minimi quadrati (LS) e il caso più semplice della regressione lineare. Applicazione del minimo chi-quadro per istogrammi binnati in approssimazione gaussiana; caso in cui deve essere applicato un modello poissoniano (quando il numero di entrate per bin è piccolo); metodo LS modificato. Introduzione al problema della combinazione di due (o più) misure; fit simultanei e regioni di controllo; media ponderata mediante l'applicazione del minimo chi-quadro nell'approssimazione gaussiana e assumendo l'assenza di correlazione. Estensione al caso correlato. Test di bontà dell'adattamento e p-valore come livello di significatività osservato. Significatività statistica di un segnale osservato (caso di variabili di Poisson); test del chi-quadro di Pearson. Approccio alternativo con un rapporto di verosimiglianza come statistica di prova nella ricerca di un nuovo segnale; esempio di un semplice esperimento di conteggio degli eventi. Lemma di Neyman-Pearson. Teorema di Wilks e sua applicabilità; esempio di rapporto di verosimiglianza considerando le ipotesi *background-only* e *signal-plus-background*. Validità delle formule asintotiche di Cowan e sua verifica con il metodo degli pseudo-esperimenti (*toys*). Livello di significatività statistica locale e osservazione (scoperta) o evidenza di un segnale. <u>Mod. F</u> Introduzione agli intervalli di confidenza classici (IC). IC di Neyman: costruzione della fascia di confidenza e sua inversione. Livello di confidenza come copertura di probabilità. IC per uno stimatore a distribuzione gaussiana. IC per la media della distribuzione di Poisson. Intervalli di confidenza utilizzando la funzione di verosimiglianza o il

	chi-quadro nel limite di grandi campioni. Intervalli binomiali e applicazione del metodo di Clopper-Pearson. Segnali rari e introduzione ai limiti superiori frequentisti (UL); esempio di esperimento di conteggio. Approccio unificato di Feldman-Cousins: come evitare il problema del flip-flopping e garantire la copertura corretta; passaggio dagli intervalli centrali agli UL. Il problema della dipendenza dei limiti superiori dalla quantità di fondo attesa in caso di eventi a segnale nullo. Approccio frequentista modificato per gli UL: il metodo CLs con esempi. Come leggere un *Brazilian Plot*. Limiti superiori utilizzando la *Profile Likelihood*. Significatività statistica globale e diluizione nota come Look-Elsewhere-Effect (LEE). Metodo dei fattori di prova; validità dell'approssimazione di Gross-Vitells e sua verifica con una tecnica di scansione basata sugli pseudo-esperimenti (Monte Carlo toys); esempio di calcolo semplificato di un p-value globale.
Testi di riferimento	Per quanto riguarda i libri di testo di riferimento e il materiale di documentazione, il materiale fornito dal docente copre l'intero corso; consiste in un insieme di diversi file pdf (in formato slides) inseriti in una pagina web dedicata al corso. Questa pagina web contiene anche utili link a preziosa documentazione online selezionata dal docente e liberamente disponibile su Internet. Come ulteriore riferimento lo studente può prendere in considerazione i libri di riferimento suggeriti nel seguente elenco e utilizzati come guida nella progettazione e nell'impostazione del presente corso: 1. G. Cowan, Statistical Data Analysis, 1998, Clarendon Press; disponibile in biblioteca e anche in formato cartaceo su richiesta del docente. 2. L. Lista, Statistical methods for Data Analysis in Particle Physics, 2020 (2a o 3a edizione), Springer Verlag, disponibile in biblioteca e anche in formato cartaceo su richiesta del docente. 3. W. J. Metzger, Statistical Methods in Data Analysis, 2010 (2a edizione), per la Radboud University di Nijmegen; gratuito online. 4. F. James, Statistical methods in Experimental Physics, 2006 (2a edizione), World Scientific; libro cartaceo fornito dal docente su richiesta.
Note ai testi di riferimento	I libri sono ovviamente più ricchi del contenuto del corso. Possono essere utilizzati come guida di riferimento per i concetti e i metodi introdotti e discussi nel corso. A volte consentono ulteriori approfondimenti utili per il lavoro supplementare qualora lo studente scelga la modalità d'esame che includa la presentazione orale (si veda – oltre – l'Opzione B).

	Offrono un'ampia varietà di esempi e applicazioni. Tuttavia, una parte rilevante degli esempi e delle applicazioni presentati in questo corso sono rielaborazioni originali di materiale esistente ed esperienze di ricerca diretta nel campo dell'HEP e possono essere trovati solo nel materiale fornito dal docente.
Materiali didattici	https://web2.ba.infn.it/~pompili/teaching.html

Valutazione	
Modalità di verifica dell'apprendimento	La scelta tra le due opzioni seguenti è lasciata a ogni studente: Opzione A - Esame orale (100%) ... oppure Opzione B - Esame orale (50%), - Foglio di lavoro supplementare con seminario finale (50%) [può essere preparato individualmente o in gruppo, ma in quest'ultimo caso i contributi individuali devono apparire chiaramente)].
Criteri di valutazione	La valutazione si basa sull'aspettativa che lo studente, studiando, abbia appreso e fatto propri, con più o meno successo, gli elementi seguenti: - la conoscenza concettuale e la comprensione dei concetti e delle metodologie statistiche avanzate utili per l'analisi dei dati sperimentali; - la consapevolezza delle caratteristiche e delle possibili problematiche di un corretto trattamento statistico nell'analisi dei *big data*; - la capacità di progettare e impostare l'analisi dei dati sperimentali; - la conoscenza e la comprensione dell'interpretazione statisticamente appropriata dei risultati dell'analisi, comprese le incertezze statistiche e sistematiche, le approssimazioni coinvolte, i possibili miglioramenti.
Criteri di misurazione dell'apprendimento e di attribuzione del voto finale	L'esame finale si svolge come nelle opzioni A o B, a scelta dello studente. Criteri di valutazione variano leggermente in base all'opzione scelta: - Conoscenza e comprensione: 40% (A) - 30% (B) - Applicazione delle conoscenze e della comprensione: 30% (A) - 20% (B) - Autonomia di giudizio: 10% (A) - 10% (B) - Abilità nel comunicare conoscenza e comprensione: 10% (A) - 20% (B) - Capacità di continuare ad apprendere: 10%(A) - 20% (B)
Altro	
	I concetti e i metodi di *Machine/Deep Learning* sono introdotti breve-mente in una sorta di panoramica a scopo di completezza. Tuttavia, si presume implicitamente che per gli studenti interessati saranno previsti corsi dedicati nel corso di laurea magistrale o, successivamente, di dottorato. Questo corso fornisce tutte le basi di conoscenza necessarie per affrontare questo tipo di corsi dedicati.