Ciências ULisboa | Estatística e Investigação Operacional

PROGRAMA DE SEMINÁRIOS 2024-2025

MESTRADO EM BIOESTATÍSTICA

DISCIPLINA: SEMINÁRIO DE BIOESTATÍSTICA, 2.º ANO – 1.º SEMESTRE

SALA 6.4.29 – SAS LAB

TRANSMISSÃO EM DIRETO VIA ZOOM: https://videoconf-colibri.zoom.us/j/93591772487

14 DE FEVEREIRO DE 2025

9:10 – ABERTURA DA SESSÃO

9:15 – 9:45 Gil Santos
A Hidden Markov Model approach to analyze fish behavior

9:45 – 10:15 Carolina Silva
The impact of AI-Assisted Decision-Making on patient satisfaction: A BREAST-Q study from the CINDERELLA trial

10:15 – 10:45 Ana Sofia Barata
Classifying app usage data with Latent Mixture Models

10:45 – 11:15 Rúben Campelo
Adherence to biological Disease-Modifying Anti-Rheumatic Drugs and the risk of development of secondary immune-mediated inflammatory diseases among rheumatoid arthritis patients: A study using administrative data

INTERVALO

11:30 – 12:00 Gonçalo Ferreira
Unraveling pancreatic cancer biology using Multi-Omics integration techniques

12:00 – 12:30 Ana Raquel Oliveira

Survival analysis in stage II colon carcinoma

12:30 – 13:00 Rita Rosado

Data analyses and integration in obstructive sleep apnea syndrome - biomarkers discovery

ALMOÇO

14:00 – 14:30 Carlos Gonçalves

Temporal changes in shark diversity and abundance in biodiversity hotspots off Cabo Verde and Eastern Australia

14:30 – 15:00 Ekaterina Nikitina

A comprehensive evaluation of missing data imputation methods in medical research using NHANES

15:00 – 15:30 Maria Leonor Chaves

Applying multivariate analytical methods to investigate Malaria

15:30 – 16:00 Nuno Domingues

Statistical methodologies for time-resolved untargeted metabolomics data: timecourse grapevine-pathogen infection datasets as case studies

INTERVALO

16:15 – 16:45 Rafaela Rodrigues

Zero-Inflated Generalized Poisson GARCH regression models for describing Pseudo-nitzschia in Lisbon Bay

16:45 – 17:15 Bianca Gasparini

Exploring phase-I clinical trial methods: SAS, R, and the path to replication

17:15 – 17:45 Gabriela Quintais

Statistical tests in ecology: context and methodology

17:45 – *FECHO DA SESSÃO*

Resumos disponíveis nas páginas que se seguem.

SEMINÁRIO de BIOESTATÍSTICA

RESUMOS DOS SEMINÁRIOS A APRESENTAR PELOS ALUNOS

MESTRADO EM BIOESTATÍSTICA, 2024-2025

14 de fevereiro de 2025

9:15 – 9:45

A Hidden Markov Model approach to analyze fish behavior

Gil Bartolomeu Santos
Orientadores: Rui Martins (DEIO-FCUL) e Fernando Sequeira (DEIO-FCUL e Escola Náutica)

Resumo: Hidden Markov Models (HMM) are widely applied in analysing animal movement, enabling the identification of distinct, often hidden, behavioural patterns based on the observed movement data. It is a probabilistic model that describes sequences of observed events influenced by underlying, unobservable factors. In this class of models, the observable event is referred to as an "observation", while the hidden, unobservable factor, is called a "state". An HMM integrates two stochastic processes: one composed by an underlying, hidden parameter process, structured as a Markov chain, and an observable process that generates the observed events, with the probabilities determined by the current hidden state. Given a dataset where observed characteristics such as temperature, pressure and acceleration coordinates were recorded from four Wels catfish in 40 second intervals for several months, this master thesis aims to develop a model that infers the fish activity (e.g., "resting", "swimming", "attacking") at each moment and estimates the transition probability between activities.

9:45 – 10:15

The impact of AI-Assisted Decision-Making on patient satisfaction: A BREAST-Q study from the CINDERELLA trial

Carolina Silva
Orientadores: Marília Antunes (DEIO-FCUL) e Giovani Silva (IST-ULisboa)

Resumo: When confronted with breast cancer, women's main concern is survival and often, little attention is given to the aesthetic outcome of the surgery to remove the tumour. On their turn, hospitals and clinical centers have limited ways of instructing the patient about breast reconstruction surgical procedures. Consequently, patients are often disappointed by their appearance once they feel their health is no longer at risk. The Cinderella project aims to improve patient's satisfaction using an APP that makes use of artificial intelligence tools to assist patients choosing the reconstruction surgery, while instructing them on what would be a poor, a fair or a good aesthetic outcome. In this project, the BREAST-Q survey, a questionnaire to assess satisfaction and quality of life for breast surgery patients, will be used to compare the results of the two arms of the trial – the intervention arm (patients who will use the APP) and the control arm (patients that will be instructed by the conventional methods). This presentation will start with a short introduction to the project, followed by a description of the data collection, the BREAST-Q survey, the statistical methods that will be used and a summary of the work done so far.

10:15 – 10:45

Classifying app usage data with latent mixture models

Ana Sofia Barata

Orientadores: Marília Antunes (DEIO-FCUL) e Giovani Silva (IST-ULisboa)

Resumo: The CINDERELLA project intends to validate an AI cloud-based approach using a web platform and a mobile application (CINDERELLA APProach) to support shared decision making in locoregional treatment. Analysing app usage patterns is relevant for understanding user behaviour. However, recorded usage time may not always reflect actual engagement. Users may leave the app open, leading to inflated usage times (outliers). This MSc project aims to apply and extend a two-component mixture modeling approaches to classify app usage time into exact and inflated. Mixture models, especially latent mixture models with covariates, offer a robust framework for addressing classification challenges when the underlying process is not fully observable. Various estimation techniques, including the Expectation-Maximization algorithm and Bayesian methods, will be tested, validated, compared, and interpreted in the context of app usage behaviour. This seminar will introduce the CINDERELLA project and the issue with the usage time data, review outlier detection and latent mixture models, present the proposed methodology, and outline initial challenges.

10:45 – 11:15

Adherence to biological Disease-Modifying Anti-Rheumatic Drugs and the risk of development of secondary immune-mediated inflammatory diseases among rheumatoid arthritis patients: A study using administrative data

Rúben Campelo

Orientadores: Eunice Carrasquinha (DEIO-FCUL) e Ersilia Lucenteforte (Univ. degli Studi di Firenze)

Resumo: Medication adherence is a key factor in patient prognosis. Common estimation measures used in healthcare databases, such as the Proportion of Days Covered and Medication Possession Ratio, often oversimplify longitudinal data, resulting in information loss. This project will utilise data from the VALORE database, including demographical and clinical data on Rheumatoid Arthritis (RA) patients treated with biological Disease-Modifying Anti-Rheumatic Drugs. The aim is to define and characterise adherence trajectories by computing adherence measures and applying 24 statistical measures to assess variability. Principal Component Analysis will identify measures explaining the greatest proportion of variability, followed by k-means clustering to group patients with similar adherence patterns. Considering the higher risk of secondary Immune-Mediated Inflammatory Diseases (IMIDs) in RA patients and the limited research on how adherence influences IMID development, this study will explore this association. Survival curves will be estimated using the Kaplan-Meier estimator, with curve comparisons via the log-rank test. Finally, Cox Regression will determine hazard ratios for IMID development.

11:30 – 12:00

Unraveling pancreatic cancer biology using Multi-Omics integration techniques

Gonçalo Ferreira

Orientadores: Eunice Carrasquinha (DEIO-FCUL) e Marta Lopes (NOVA FCT)

Resumo: Pancreatic cancer, characterized by its high lethality, aggressive nature, and potential for rapid metastasis, poses significant challenges for early detection and treatment. This study will leverage clinical, genomic and transcriptomic data from The Cancer Genome Atlas (TCGA) to perform sub-type clustering, sub-type classification and identify novel biomarkers. The project's methodology

will involve a detailed examination of each dataset independently, using statistical approaches such as independence tests, t-tests, correlation analysis, PCA, clustering algorithms, logistic regression, and survival analysis. Following this exploratory phase, three advanced machine learning techniques — Multi-Omics Factor Analysis (MOFA+), iClusterBayes, and Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO) — will be employed for data integration. By integrating these diverse data types, the study aims to provide valuable insights into the molecular heterogeneity of pancreatic cancer, facilitating subtype classification, improving patient stratification, and contributing to the development of personalized treatment strategies. Ultimately, the project aims to explore the complexities of pancreatic cancer through the integration of multi-omics data.

12:00 – 12:30

Survival analysis in stage II colon carcinoma

Ana Raquel Oliveira
Orientadores: Tiago Domingues (DEIO-FCUL) e José Passos Coelho (Hospital da Luz Lisboa)

Resumo: Colon carcinoma remains a significant global health challenge, being the third most common cancer worldwide and the second leading cause of cancer-related mortality (1). A retrospective cohort of colon cancer patients that underwent surgery between 2012-2018 is analyzed to identify significant prognostic factors, using R software. The aim is to characterize the study population and assess the survival of these patients through adjuvant chemotherapy, using the Kaplan-Meier non-parametric estimator. Logistic regression models will be used to identify risk factors for recurrence and death, as well as Cox regression to identify factors that influence the survival. If the assumption of proportional hazard does not verify, other statistical methods will be used to address this problem, such as the Adaptive Neyman's Smooth tests (2). By delineating the contribution of observed variables to overall and disease-free survival, this research aims to enhance predictive accuracy and provide insights for individualized treatment approaches, contributing to improved patient outcomes and resource allocation in colon cancer management.

12:30 – 13:00

Data analyses and integration in obstructive sleep apnea syndrome - biomarkers discovery

Rita Rosado
Orientadores: Marília Antunes (DEIO-FCUL) e Deborah Penque (INSA)

Resumo: Obstructive Sleep Apnea Syndrome (OSA) is a common sleep disorder characterized by repetitive upper airway obstruction during sleep, leading to fragmented sleep and intermittent hypoxia. This condition is associated with increased risks of cardiovascular, metabolic, and neurocognitive disorders, underscoring the need for improved diagnostic and treatment strategies. Recent advances in proteomics and metabolomics have provided a wealth of data, offering new opportunities to uncover biomarkers that can enhance understanding of OSA and predict treatment responses. The Laboratory of Proteomics at INSA has generated comprehensive proteomic and metabolomic datasets, complemented by clinical and biochemical profiles of a group of patients with OSA, before and after six months with Positive Airway Pressure (PAP) treatment, and Snorers subjects as controls. This dataset will be used to employ and test different computational and statistical approaches, and the main goal of this thesis is to uncover candidate proteomic and metabolomic biomarkers as diagnostic and/or therapeutic monitoring tools for OSA. In this seminar we will talk about the dataset, type of variables, methods that will be used in this study and some preliminary results.

14:00 – 14:30

Temporal changes in shark diversity and abundance in biodiversity hotspots off Cabo Verde and Eastern Australia

Carlos Jr. Gonçalves
Orientadores: Rui Rosa (MARE, ARNET e FCUL) e Tiago Marques (DBA-FCUL, Univ. St Andrews)

Seminar presented in Portuguese.

Resumo: Understanding temporal changes in shark diversity and abundance in biodiversity hotspots is fundamental to understanding population patterns and conservation. This study investigates the influence of environmental factors on shark populations at Boavista Island (Cape Verde) and Lizard Island (Great Barrier Reef, Australia). The first chapter focuses on the study of juvenile nurse sharks (Rhizoprionodon acutus) in a nursery area in Cape Verde. The influence of water temperature, salinity, chlorophyll-a concentration and dissolved oxygen on abundance will be evaluated at multi-annual, annual and monthly scales. The preliminary analysis shows the existence of seasonality in the species' abundance. The second chapter will examine decadal changes in shark abundance in a diverse cast of species such as tiger sharks, lemon sharks and grey nurse sharks, in order to relate changes to environmental conditions. The results will reinforce the importance of environmental factors and hotspots for global shark conservation.

14:30 – 15:00

A comprehensive evaluation of missing data imputation methods in medical research using NHANES

Ekaterina Nikitina
Orientadores: Helena Mouriño (DEIO-FCUL)

Resumo: Missing data is a common issue in medical research, leading to bias and reduced statistical power. This study evaluates missing data imputation methods using NHANES, focusing on three mechanisms: MCAR (Missing Completely at Random), MAR (Missing at Random), and NMAR (Not Missing at Random). The research follows two steps: (1) controlled experiments introducing artificial missingness to compare imputation accuracy using Mean Squared Error (MSE) and F1-score, and (2) real-world analysis applying these methods to NHANES with actual missing values. Various techniques, including single imputation, multiple imputation, and model-based approaches, are assessed for statistical accuracy and predictive performance. Results highlight the strengths and limitations of different methods, offering recommendations for handling missing data in medical research. The study contributes to improving data integrity and ensuring robust statistical analysis in epidemiological studies.

15:00 – 15:30

Applying multivariate analytical methods to investigate Malaria

Maria Leonor Chaves
Orientadores: Eunice Carrasquinha (DEIO-FCUL) e Ângelo Chora (GIMM)

Seminar presented in Portuguese.

Resumo: Malaria, caused by Plasmodium parasites, remains a severe global health issue, affecting millions and causing thousands of deaths annually. Infection in the mammalian host begins with the bite of an infected mosquito, after which parasites establish in the liver. There, parasites undergo

extensive replication within hepatocytes before entering the bloodstream, where they cyclically infect red blood cells. In this study, host metabolic alterations during the hepatic phase of the infection will be identified using multivariate statistical techniques on an untargeted metabolomics dataset. These include Principal Component Analysis to reduce data complexity, and Cluster Analysis to categorize observations by similarities and dissimilarities. Normality conditions will be verified applying parametric tests (t-test, ANOVA, ANCOVA) and, if unmet, non-parametric tests ($\chi 2$ test, Wilcoxon, Kruskal-Wallis, Mann-Whitney, Spearman's Rank). To address outliers' mitigation, robust statistical methods will be explored, including Robust Principal Component Analysis. Additionally, the study will discuss the most suitable methodology, considering the required information and objectives.

15:30 – 16:00

Statistical methodologies for time-resolved untargeted metabolomics data: timecourse grapevine-pathogen infection datasets as case studies

Nuno Domingues
Orientadores: Lisete Sousa (DEIO-FCUL) e Marisa Maia (GPS Lab, BioISI e DBV-FCUL)

Seminar presented in Portuguese.

Resumo: Recent years have been marked by the advancement of high-throughput metabolomics technologies. Biological research has reflected such advancements into complex datasets with high dimensionality, noise, and missing values. Moreover, experimental designs are becoming increasingly complex, often involving multiple factors and measurements over time. Methodologies commonly used to analyse metabolomics data generally fail to fully incorporate the underlying experimental design in the analysis, such as neglecting time as a factor, which can result in overlooking dynamic insights into biological processes. This presentation will focus on ANOVA Simultaneous Component Analysis (ASCA), which combines ANOVA with dimensionality reduction to handle high-dimensional data while modelling experimental designs. Thus, providing a robust framework for analysing complex datasets, enabling a deeper understanding of molecular mechanisms in life processes. Other methodologies to tackle the challenges posed by metabolomics data will also be briefly discussed, such as imputation of missing values. These methodologies will be studied in the context of plant metabolomics, particularly plant-pathogen interaction.

16:15 – 16:45

Zero-Inflated Generalized Poisson GARCH regression models for describing Pseudo-nitzschia in Lisbon Bay

Rafaela Lopes Rodrigues
Orientadores: Helena Mouriño (DEIO-FCUL) e Valdério Reisen (Univ. Federal do Espírito Santo)

Resumo: Phytoplankton play a vital role in aquatic ecosystems, serving as a primary food source for various marine organisms. However, excessive nutrient availability can result in harmful algal blooms, which produce toxic compounds that affect marine birds, mammals, and humans. Among these, diatoms of the genus *Pseudo-nitzschia* are particularly concerning due to their ability to produce domoic acid, a potent neurotoxin responsible for Amnesic Shellfish Poisoning outbreaks worldwide. Therefore, understanding the seasonal and spatial dynamics of Pseudonitzschia blooms is crucial for assessing their ecological and health impacts. This study analyzes weekly water samples collected from June 2001 to May 2005 in Cascais (Lisbon Bay) to determine *Pseudo-nitzschia spp*. concentrations. While previous research applied a Zero-Inflated Generalized Poisson Regression

Model to describe Pseudo-nitzschia variability and the influence of environmental factors such as sea surface temperature and upwelling indices, this study extends the analysis by incorporating Generalized Autoregressive Conditionally Heteroscedastic (GARCH) models to capture time-series volatility.

16:45 – 17:15

Exploring phase-I clinical trial methods: SAS, R, and the path to replication

Bianca Gasparini
Orientadores: Marília Antunes (DEIO-FCUL)

Resumo: Phase I clinical trials primarily assess a drug's safety, tolerability, and pharmacokinetics rather than its efficacy, which is the focus of later phases. Statistical methods in this phase include bioavailability and bioequivalence testing, fixed and mixed effects ANOVA models, sample size and power calculations. Historically, SAS has been the industry standard for regulatory drug applications due to its reliability. However, its slow adaptation to new statistical methods has increased interest in R, an open-source alternative with faster implementation of new techniques via package updates. As pharmaceutical companies and regulatory agencies explore transitioning from SAS to R, ensuring consistency and accuracy in numerical algorithms is crucial. This seminar focuses on explaining the statistical methodologies used in these trials and demonstrating their implementation in SAS and R. Using generated datasets, key functions and coding approaches in both platforms will be presented, highlighting how R can replicate SAS-based analyses. This initial exploration sets the foundation for a future comparison of results across both platforms.

17:15 – 17:45

Statistical tests in ecology: context and methodology

Gabriela Xavier-Quintais
Orientadores: Tiago Marques (DBA-FCUL, U. St Andrews) e Daniël Lakens (U. Technology, Eindhoven)

Resumo: The use of hypothesis testing and p-values has been ongoing since the 18th century, well before Fisher formally defined the concept in 1925. Over this extended debate scientists have identified more than 6 paradoxes or fallacies related to p-values, over 12 common misinterpretations, and diagnosed several limitations of hypothesis testing. In ecology, some of these issues seem particularly relevant, due to inherent complexity and heterogeneity of ecosystems, high environmental variability, and complex and dependent interactions between biotic and abiotic factors. Despite an abundance of statistical and theoretical criticism on hypothesis testing, it has not been examined to what extent hypothesis testing is suitable for the scientific questions asked in ecology in practice. We will examine how hypothesis testing is used in scientific practice by meta-scientifically evaluating a random sample of 120 papers in 12 general ecology journals, and evaluating their suitability to answer questions ecologists are interested in.