5th Young-ISA Meeting

Transferable Skills: Insights from Academia and Industry

December 9, 2024

Lynch Theatre, O'Brien Centre for Science

University College Dublin

09:00-09:30	Registration	
09:00-09:50	Welcome address and dedication: Kathleen O'Sullivan's contribution to the Young-ISA	
Keynote talk 1 (Chair: Shirin Moghaddam)		
09:50-10:30	Nial Friel	Navigating a career in statistics: embracing uncertainty!
Session 1 (Chair: Gabriel Palma)		
10:30-10:45	Amanda Forde	Statistical corrections for Winner's Curse bias in genetic association studies
10:45-11:00	Pedro Menezes De Araújo	Summarising mortality data with time-dependent beta latent variable models
11:00-11:15	Megan O'Sullivan	Using Simulation to increase diversity and success rates in respiratory clinical trials
11:15-11:45	Coffee break	
Keynote talk 2 (Chair: Nastaran Sharifian)		
11:45-12:25	Owen McGrath	From Wrenches to Regression: Building a Career in Statistics
Session 2 (Chair: Silvia D'Angelo)		
12:25-12:40	Ultán Doherty	Model-Based Clustering with Sequential Identification of an Unspecified Number of Outliers
12:40-12:55	Conor Hackett	A Method to Identify Wildfire Ignition Points and Propagation Durations Using Genetic Algorithms
12:55-13:10	Kate O'Donovan	Adaptive Mesh Construction for the Numerical Solution of Stochastic Differential Equations with Markovian Switching
13:10-14:10	Lunch	

Keynote talk 3 (Chair: Luke Kelly)		
14:10-14:50	Órlaith Burke	Research, Growth, & Leadership: Continuous Skills for an Evolving Career
Round table (Chair: Szymon Urbas, Thais Pacheco Menezes)		
14:50-15:25	Nial Friel, Owen McGrath, and Órlaith Burke	
15:25-16:10	Coffee break and poster session	
Joint Young-ISA and BIR Session (Chair: Catherine Higgins)		
16:10-16:25	Eva Ryan	Is the relationship between chronic pain and mortality causal?
16:25-16:40	Clara Panchaud	Incorporating Memory into Spatially-Explicit Capture-Recapture Models
16:40-16:55	Gabriel Palma	Forecasting insect abundance using time series embedding and machine learning
16:55-17:00	Event closing and announcement of poster winners	

Navigating a career in statistics: embracing uncertainty!

Keynote speaker: Nial Friel

University College Dublin

A career in statistics is both varied and rewarding. It is a unique feature of our profession that we can collaborate with scientists, engineers, economists, effectively anyone who generates data and that we can "play in everyone's backyard". This talk will give a personal reflection on a career as an academic statistician. I'll present some lessons learned along the way as well as some pointers for what a career in statistics may look like in the age of AI.

Statistical corrections for Winner's Curse bias in genetic association studies

Amanda Forde

University of Galway

Genetic variants are differences in the DNA sequence between individuals. A genetic association study is a research approach that compares the DNA sequences of a large group of people in order to identify genetic variants statistically associated with a specific complex trait or disease. It is often observed that the estimated variant-trait associations tend to be smaller in magnitude in follow-up replication studies compared to the initial discovery study. This is largely due to a phenomenon known as *Winner's Curse*, which causes an upward bias in association estimates of significant variants in the discovery study.

Correcting for this bias is challenging, but several statistical methods have been proposed to improve the accuracy of association estimates derived from discovery studies. In our work, we focus on reviewing methods that require only summary-level data to make adjustments. We suggest several modifications to improve these approaches and introduce a novel method based on the parametric bootstrap. Specifically, we examine the strengths and limitations of three main approaches, namely conditional likelihood, empirical Bayes and the parametric bootstrap.

Comparative method performance, in terms of both bias and mean square error, is assessed using a wide range of simulated data sets as well as real data sets from the UK Biobank. Our findings demonstrate the advantage of adjusting for *Winner's Curse* bias by using methods which jointly consider the effect sizes of all variants, as opposed to implementing corrections independently.

Additionally, we have developed an R package, namely 'winnerscurse', which enables users to easily apply *Winner's Curse* adjustment methods to summary-level data, yielding more accurate association estimates for significant genetic variants.

Summarising mortality data with time-dependent beta latent variable models

Pedro Menezes De Araújo

University College Dublin

Age-specific central mortality rates provide a snapshot of population mortality at the country level at a given point in time. Due to the high dimensionality of the data, summarizing mortality information is essential for various analyses, such as visualization and clustering. We propose the use of beta latent variable (BLV) models to summarize mortality information without data transformation. A time-dependent version of the BLV model is developed by incorporating an autoregressive prior for the latent effects. This model aims to represent mortality rate data with a small set of K latent effects while accounting for time dependence between these effects. Inference is performed using Bayesian methods, with posterior samples generated via Hamiltonian Monte Carlo. The BLV model is applied to central mortality rates from the Human Mortality Database, covering 41 countries and 24 age-specific mortality rates over several periods. The time-dependent BLV outperforms the standard Gaussian factor analysis model applied to log mortality rates and demonstrates that BLV models can effectively summarize mortality rate data.

Using Simulation to increase diversity and success rates in respiratory clinical trials

Megan O'Sullivan

Queen's University Belfast

In 2017 it was estimated that 544.9 million people worldwide had a chronic respiratory disease.Despite this high global disease burden, a number of chronic respiratory diseases such as bronchiectasis currently have no licenced treatment, in part due to a lack of successful clinical trials. To it is vital that we review past trials so as not to replicate any previous errors in design.

Trials failing for efficacy and low recruitment can both be traced back to highly restrictive eligibility criteria. These criteria are in place to give a trial the best possible chance of succeeding through reducing confounding error, and selecting patients who will maximize the observed treatment effect. However in many scenarios, the eligibility criteria severely restrict the number of patients who could potentially be enrolled in a trial. Clinical trials have a high risk of failure, and sponsors are reluctant to add new groups of patients who might introduce additional risk.

When we consider that the results from a clinical trial only represent a small subgroup of a disease population, concerns about the applicability of these results to the real world must be raised. One review that investigated the generalisablilty of bronchiectasis trials to the actual population, found that less than 10\% of the patient population were eligible to participate in some trials.

The aim of this research is to use simulation to demonstrate that patient diversity in clinical trials can be increased, and potentially contribute to reducing failure rates in clinical trials. Using data from the BRONCHUK patient registry, we will simulate past bronchiectasis trials. We will train models on clinical trial data to recreate these trials using patients who didn't originally take part. When simulating these trials, we will also include patients who weren't originally eligible to take part. However, when these patients are included and randomised to the new treatment group, we will reduce their predicted treatment effect to make it lower than in those who were originally eligible.

By simulating these trials, adding varying proportions of patients in subgroups who were not originally eligible to take part, we will be able to determine whether the addition of these patients will affect the integrity of the trial. The simulations will be evaluated by looking at how many patients the trial would need to show the minimum detectable difference between the groups, how long the trial would take and the probability of the trial succeeding.

From Wrenches to Regression: Building a Career in Statistics

Keynote speaker: Owen McGrath

University of Limerick

In this talk I will share my personal journey from working on a construction site to becoming a TU (technological university) lecturer and pursuing a PhD in statistics. Drawing on my experiences across diverse roles spanning construction, engineering, data science, and now academia, I will highlight the transferable skills that have been essential throughout my career.

This honest account of successes, challenges, and learnings aims to provide practical insights to students and early career researchers as they navigate their own pathways to industry or academia.

Model-Based Clustering with Sequential Identification of an Unspecified Number of Outliers

Ultán Doherty

Trinity College Dublin

Co-authors: Paul McNicholas, McMaster University; Arthur White, Trinity College Dublin.

The presence of outliers can prevent clustering algorithms from accurately determining an appropriate group structure within a data set. We present outlierMBC, a model-based approach for sequentially trimming outliers and clustering the remaining observations. Our method fits a Gaussian mixture model to the data and identifies outliers using the estimated Mahalanobis distance of observations within each cluster. It does not require the number of outliers to be pre-specified.

it can be difficult to classify observations as outliers without knowing the data's clusters and the presence of outliers interferes with modelling clusters correctly, we use an iterative method to identify outliers one-at-a-time, following a similar procedure to the recently developed OCLUST method (Clark & McNicholas, 2024). At each iteration, outlierMBC removes the least well-fitting observation and fits a Gaussian mixture model to the remaining data. The method continues to trim potential outliers until a pre-set maximum number of outliers is reached, then retrospectively identifies the point at which an optimal number of outliers had been removed. Since the cluster solution becomes easier to identify as outliers are removed, this should improve our ability to correctly distinguish between borderline outliers and observations.

This approach utilises the assumption that observations in each cluster follow a Gaussian distribution and the fact that the sample Mahalanobis distances of Gaussian distributed observations are Beta distributed when scaled appropriately. We choose the number of trimmed outliers which minimises a dissimilarity between this theoretical Beta distribution and the observed distribution of the scaled sample Mahalanobis distances. We also developed a secondary mechanism for determining the optimal number of outliers which compares the expected and observed numbers of extreme observations and is less sensitive to departures from Normality within the clusters. outlierMBC differs from OCLUST by using the distribution of scaled sample Mahalanobis distances instead of the distribution of scaled and shifted subset log-likelihoods, which eliminates the need for several assumptions and achieves a significant reduction in computational time. Our method performs strongly compared to other leading algorithms when applied to a range of simulated data sets and provides reasonable solutions when applied to real data sets.

A Method to Identify Wildfire Ignition Points and Propagation Durations Using Genetic Algorithms

Conor Hackett

Maynooth University

A critical research area regarding wildfire modelling that is often overlooked is the task of finding where a wildfire started and how long that wildfire burnt for. A review of the literature revealed that there are no automated methods for the detection of wildfire ignition points using wildfire burn scars. This presentation describes a novel method called the WSGA (Wildfire Source Genetic Algorithm) to find the ignition points and the propagation time of a wildfire, given the environmental condition and the burn scar. The WSGA encodes a bitstring that corresponds to regions described by polygons within a wildfire simulating program called the IGS (Irregular Grid Software). In the bitstring, the bit value specifies whether that polygon contains a wildfire ignition point and is therefore a wildfire source polygon. The WSGA also generates a value representing the propagation time of the wildfire, called the simulation duration. Multiple bitstrings with simulation durations are generated. The WSGA uses these bitstrings and simulation durations to populate a genetic algorithm. The genetic algorithm compares the WSGA created wildfires seeded with the information in the bitstrings and the simulation durations to the original burn scar. The bitstring and simulation durations of the simulated wildfires that most closely resemble the original burn scar are then identified. These are then combined, and the process continues. This gradually generates a population of bitstrings and simulation durations that produce wildfires which more closely resemble the original burn scar. To evaluate the final outputted wildfires of the WSGA, a relative distance error was calculated by summing the Euclidian distances between source polygons produced by the WSGA and the source polygons of the original burn scar relative to the diameter of the original burn scar. Depending on the scenario the WSGA had a relative distance error in the range of [0, 1.25]. A relative simulation duration error was also calculated by finding the difference between the WSGA simulation duration and the original burn scar simulation duration, relative to the original burn scar simulation duration. Depending on the scenario the relative simulation duration error had a range of [0.0006, 0.49].

Adaptive Mesh Construction for the Numerical Solution of Stochastic Differential Equations with Markovian Switching

Kate O'Donovan

University College Cork

Stochastic differential equations (SDEs) are used to model the evolution of real-world phenomena subject to random noise and uncertainty. Consider, for example, asset prices or stochastic interest rates in finance, models of ecological systems with complex interaction between species or models of chemical reactions in biological cells. The random noise may act as a diffusion, for example reflecting market volatility, or as a jump process, for example when an ecosystem is influenced by a random external event.

For most nonlinear SDEs there is no closed-form solution and typically numerical methods are used by modellers. However, standard schemes based on solving to a final time using a uniform step size are not applicable for highly nonlinear systems and the methods that do exist are often inefficient.

In this talk I discuss the use of an adaptive mesh construction strategy for SDEs which are subject to impulsive shocks at random intervals. These shocks are taken as systemic shifts modelled in SDE coefficients according to the evolution of a Markov chain.

I will motivate and characterise this strategy and provide a strong convergence analysis for the numerical method implemented on the random mesh it generates. Implementation will be illustrated via a model of telomere shortening in jackdaws.

Research, Growth, & Leadership: Continuous Skills for an Evolving Career

Keynote speaker: Órlaith Burke

Global Innovation Director - Accenture

Órlaith is an alumna of UCD, yet having spent a total of 8 years here between her BSc in Mathematical Sciences and her PhD in Statistics she still gets lost on the new campus. Chapter One of Órlaith's career was her PhD (completed via a stint as a Visiting Scholar at Columba University in NY). Órlaith spent Chapter Two of her career as a lecturer of statistics at the University of Oxford. Chapter Three is at Accenture - specifically in their global innovation centre, The Dock here in Dublin. Today, she will be sharing her reflections on the similarities of the skills she has used, reused, learned and developed across all three chapters.

Is the relationship between chronic pain and mortality causal?

Eva Ryan

University of Limerick

Co-authors: Hanna Grol-Prokopczyk, University at Buffalo; Christopher R. Dennison, University at Buffalo; Anna Zajacova, University of Western Ontario; Zachary Zimmer, Mount Saint Vincent University.

Chronic pain is a common condition among older adults that affects both physical and mental health. However, the nature of the relationship between chronic pain and mortality remains unclear. Whether chronic pain itself causes increased mortality risk, or is merely associated with increased mortality due to confounding factors, has important implications for the development of effective public health policies and clinical interventions.

As it would be unfeasible and unethical to conduct a randomised trial of chronic pain exposure, observational data must be used to investigate this relationship. In this work, we estimate the causal effect of chronic pain on mortality over a 20-year period using data from the U.S. Health and Retirement Study. A directed acyclic graph was created to illustrate the assumed causal structure of the pain-mortality relationship and to aid the identification of an appropriate confounder adjustment set. Both propensity score matching and inverse probability weighting were used to adjust for measured confounding. Cox proportional hazards models were then fitted to the matched (or weighted) datasets to estimate the effect of chronic pain on mortality as a hazard ratio.

While unadjusted analyses found a strong association between chronic pain and mortality, the estimated causal effect after adjusting for confounding was modest. In some cases, the results were also compatible with no causal effect. This attenuation suggests that confounders play a substantial role in the observed association. We also conducted an alternative analysis treating depressive symptoms as a mediator rather than a confounder in the pain-mortality relationship. This analysis found stronger evidence of a modest effect of chronic pain on mortality, highlighting the sensitivity of the results to assumptions about the underlying causal structure.

Incorporating Memory into Spatially-Explicit Capture-Recapture Models

Clara Panchaud

University of Edinburgh

Estimating wildlife species abundance and distribution underpins the conservation and management of animal populations and natural reserves. Spatially explicit capture-recapture (SECR) models are often applied to data collected from camera traps to obtain population and spatial density estimates of animal populations, which are important to assess the conservation status and the impact of conservation actions. SCR models include spatial correlation by assuming that animals are more likely to be observed by sensors close to their activity centre. However, these models consider that an individual's known location from a previous trap sighting does not influence the probability of being seen at future times at the given trap locations, which is unrealistic as animals move through space and time smoothly. We propose a new continuous-time modeling framework to account for spatial correlation of observations due to both an individual's (latent) activity centre and (known) observed locations from previous captures. I'll present the novel model as well as simulations and a case study.

Forecasting insect abundance using time series embedding and machine learning

Gabriel Palma

Maynooth University

Implementing insect monitoring systems provides an excellent opportunity to create accurate interventions for insect control. However, selecting the appropriate time for an intervention is still an open question due to the inherent difficulty of implementing on-site monitoring in real-time. This decision is even more critical with insect species that can abruptly increase population size. A possible solution to enhance decision-making is to apply forecasting methods to predict insect abundance. However, another layer of complexity is added when other covariates are considered in the forecasting, such as climate time series collected along the monitoring system. Multiple possible combinations of climate time series and their lags can be used to build a forecasting method. Therefore, this research paper proposes a new approach to address this problem by combining statistics, machine learning, and time series embedding. We used two datasets containing a time series of aphids and climate data collected weekly in Coxilha and Passo Fundo municipalities in Southern Brazil for eight years. We conduct a simulation study based on a probabilistic autoregressive model with exogenous time series based on Poisson and negative binomial distributions to check the influence of incorporating climate time series on the performance of our approach. We pre-processed the data using our newly proposed approach and more straightforward approaches commonly used to train machine learning algorithms in time series problems. We evaluate the performance of the selected machine algorithms by looking at the Root Mean Squared Error obtained using one-step-ahead forecasting. Based on Random Forests, Lasso-regularised linear regression, and LightGBM regression algorithms, our novel approach yields competitive forecasts while automatically selecting insect abundances, climate time series and their lags to aid forecasting.

Acknowledgements

We dedicate this event to Kathleen O'Sullivan, who was instrumental in the creation of the Young-ISA, and was a treasured colleague and friend to many of us in the statistical community.

The Young-ISA would like to express our sincere thanks to everyone who contributed talks and posters to this event, and to all who participated in the workshop. We thank the Irish Statistical Association for funding this workshop through their short course funding scheme. Thank you to the British and Irish Region (BIR) of the International Biometric Society (IBS) for cooperation and financial contribution toward organising our joint Young-ISA and BIR session. We would especially like to give thanks to the school office in the School of Mathematics and Statistics at University College Dublin for their help in organizing this event. Thank you also to Nial Friel, Head of School, for facilitating the event.

Organising committee

University College Dublin: Catherine Higgins, Kate Finucane, and Thais Pacheco Menezes.

Young-ISA: Shirin Moghaddam, Gabriel Rodriques Palma, Nastaran Sharifian, Silvia D'Angelo, Autumn O'Donnell, Emily Gribbin, Owen McGrath, Luke Kelly, Szymon Urbas, Fatima-Zahra Jaouimaa, and James Sweeney.