Manjari Das *Curriculum Vitae*

CONTACT INFORMATION

E-mail: manjari8d@gmail.com ***Mobile:*** +1-6518085546
Github: github.com/mqnjqrid ***Website:*** mqnjqrid.github.io

RESEARCH INTERESTS

Primary: Missing data problems, Nonparametric Statistics, Causal Inference, Stochastic Process, Network Models.
Secondary: Statistical Computing, Bayesian modelling, Convolutional Neural Networks.

EDUCATION

Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213.

Ph.D. in Statistics and Data Science, August 2016-May 2022 QPA 3.95.
Advisor: Edward H. Kennedy (edward@stat.cmu.edu). (*cumulative*)
Thesis Committee: Larry Wasserman (larrywasserman.cool@gmail.com)
Robin Mejia (rmejia@andrew.cmu.edu)
Sivaraman Balakrishnan (siva@stat.cmu.edu)
External: Nicholas Jewell (Nicholas.Jewell@lshtm.ac.uk), Chair, Biostatistics & Epidemiology, London School of Hygiene and Tropical Medicine

Indian Statistical Institute (ISI), Kolkata, West Bengal, India.

Masters of Statistics, July 2014-June 2016 First Division (distinction).

Bachelor of Statistics (Honours), July 2011 *to* June 2014 First Division (distinction).

AWARDS AND OPPORTUNITIES

- **Dietrich College Teaching Fellowship** for Spring 2021.
- Conference travel award from Graduate Student Assembly in Summer 2019.
- Conference travel award for International Conference On Statistical Distributions and Applications.
- Mukul Chaudhuri Memorial Prize from the *President of India* at the 48^{th} Convocation of Indian Statistical Institute, for *"the best girl student in first year of bachelor's program in academics"*.
- **INMO** (Indian National Mathematics Olympiad) Awardee 2011. All India rank 6.
  Opportunity to attend one month long ***International Mathematics Olympiad Training Camp*** *at Homi Bhabha Centre for Science Education, Mumbai in 2011.*
- **INMO** Certificate Of Merit in 2010.
- Recipient of **NBHM** (National Board for Higher Mathematics) Scholarship July 2011 to June 2014.
- All India Rank 8 in **Kendriya Vidyalaya Sangathan Junior Mathematics Olympiad** 2008. Opportunity to attend two one-week long math training camps at Bangalore, 2009 and Mumbai, 2010 organised by the **NBHM** (National Board for Higher Mathematics).

EXPERIENCE

Dissertation Projects

▲ **Title**: **Doubly robust capture-recapture methods for estimating population size**. *Manuscript under review arxiv:2104.1409*

Co-authors: Edward H. Kennedy, Associate Professor, Department of Statistics and Data Science, Carnegie Mellon University, and Nicholas P. Jewell, Chair, Biostatistics & Epidemiology, London School of Hygiene and Tropical Medicine

Abstract: Estimation of total population size using incomplete lists has long been an important problem across many biological and social sciences. For example, partial and overlapping lists of casualties in the Syrian civil war are constructed by multiple organizations, and it is of great interest to use this information to estimate the magnitude of destruction of the war. Earlier approaches to solving these kinds of problem have either used strong parametric assumptions or suboptimal plugin-type nonparametric techniques; however, both

approaches can lead to substantial bias, the former via model misspecification and the later via smoothing. Under an identifying assumption that two lists are conditionally independent given covariate information, we make the following advances: First we derive a nonparametric efficiency bound for estimating the capture probability, based on the efficient influence function and we have proved double-robustness for the remainder term. Then we construct a bias-corrected estimator that attains this bound under weak nonparametric conditions. Finally, finite-sample properties of the proposed estimator are studied with simulations, and we apply our methods on Peruvian armed conflict killings dataset from 1980-2000.

▲ **Title**: **drpop: Efficient and Doubly Robust Population Size Estimation in R**. *Submitted arxiv:2111.07109*

Co-author: Edward H. Kennedy, Associate Professor, Department of Statistics and Data Science, Carnegie Mellon University.

Abstract: This paper introduces the R package drpop to flexibly estimate total population size from incomplete lists. Total population estimation, also called capture-recapture, is an important problem in many biological and social sciences. A typical dataset consists of incomplete lists of individuals from the population of interest along with some covariate information. The goal is to estimate the number of unobserved individuals and equivalently, the total population size. drpop flexibly models heterogeneity using the covariate information, under the assumption that two lists are conditionally independent given covariates. This can be a much weaker assumption than full marginal independence often required by classical methods. Moreover, it can incorporate complex and high dimensional covariates, and does not require parametric models like other popular methods. In particular, our estimator is doubly robust and has fast convergence rates even under flexible non-parametric set-ups. drpop provides the user with the flexibility to choose the model for estimation of intermediate parameters and returns the estimated population size, confidence interval and some other related quantities. In this paper, we illustrate the applications of drpop in different scenarios and we also present some performance summaries.

▲ **Title**: **Population size estimation in the capture-recapture set-up under partial identification**. *Working paper*

Co-author: Edward H. Kennedy, Associate Professor, Department of Statistics and Data Science, Carnegie Mellon University.

Abstract: Capture-recapture problems require additional assumptions to ensure parameter identifiability. Existing literature has used some lack of dependence assumption among the lists which may or may not be valid. We look into a weaker assumption that ensures only partial identifiability. We derive sharp bounds on the total population size in a sensitivity analysis set-up using a weak margin assumption. Instead of point estimates we present confidence bounds for the total population size. We further present the finite-sample error in the coverage guarantee of the estimated confidence interval.

▲ **Title**: **Conditional estimation of population size in the capture-recapture set-up**. *Working paper*

Co-author: Edward H. Kennedy, Associate Professor, Department of Statistics and Data Science, Carnegie Mellon University.

Abstract: Most of the existing literature in the area of total population size estimation focus on a single estimate. But in the presence of continuous covariates, it is often of interest to estimate the population size as a function of that covariate, for example, time or age of individuals in the population. This problem is similar to the case where we are interested in the whole population size. However, because we are now conditioning on some continuous covariate, this makes the problem more complicated. Moreover, the number of quantities to be estimated is larger while the available training data is limited. Also, oracle efficiency rates are not guaranteed.

Other Projects

- **Title**: **Linking galaxies across time using morphological statistics on the Illustris data**.

  Supervisors: Ann Lee, Professor and Peter Freeman, Professor, Carnegie Mellon University, Department of Statistics and Data Science, and Gregory F Snyder, Astronomical Data Scientist, Space Telescope Science Institute.

  Abstract: The details of galaxy evolution physics are complex and are linked to the global properties of the Universe. In this project, we aim to link galaxies at a given time to their progenitors in the Illustris Project data. We use random forest to predict the past mass rank of a galaxy at a different time which is used to link the galaxies. We applied our model on the CANDELS data. For real data, there is no direct way to test the linking at an individual level. We tested the validity of the whole estimated and true progenitor populations with two-sample paired t-test and Kolmogorov-Smirnov test.

- **Project**: **Simulated evolution of Google+ Social Network to study relation of edge formation to common covariates of the nodes**.

  Supervisor: Diganta Mukherjee, Professor, Sampling and Official Statistics Unit, Indian Statistical Institute Kolkata.

  Abstract: This project aims to determine whether the immediate social connections of a person can explain his or her behaviour. The available data is a snapshot of a Google+ sub-network from snap.stanford.edu. We used simulation to study the evolution of the network from a randomly selected sub-graph with the same set of nodes. The evolution uses the distance between two nodes (in terms of their covariate vectors) to randomly connect them with an edge. Hence, if people with similar covariates are more likely to connect, the simulated network will be closer to the true network.

- **Title**: **Prediction of the factors for determination of box office collection of a movie in Bollywood in 2012**.

  Supervisors: Smarajit Bose, Professor, Interdisciplinary Statistical Research Unit, Indian Statistical Institute, Kolkata.

  Abstract: The data, collected by the group members, was a list of all movies released in Bollywood between September 2012 and August 2013. the goal was to identify factors which can estimate the box-office collection. The movies were classified on the basis of their popularity and revenue generated. We used imputation for missing values and random forest to identify the factors which affect the box office collection.

- **Title**: **Consumption pattern of various commodities in different states of India**.

  Supervisor: Kundu Pally, Director, National Sample Survey Organization, Kolkata, India.

  Abstract: The data is from the of 68^{th} Round of National Sample Survey (Schedule 1.0 Type II) on Consumer Expenditure. We analysed the demand pattern on necessities of different commodities from this data. We fitted a **Modified Working Leser Model** to consider budget share of particular items as a function of total expenditure of a household. We computed the **elasticity of the items** and used these to classify the items as inferior, essential and luxury.

- **Project**: **Distribution regression in semi-supervised learning framework.**

  Co-author: Kwhangho Kim, fellow PhD student, Department of Statistics and Data Science and Barnabas Poczos, Professor, Department of Machine Learning, Carnegie Mellon University.

  Abstract: 'Distribution Regression' refers to those regression problems where the response variable Y (output) depends on the covariate variable P (input) in which P is a probability

distribution. In this work, we develop an algorithm for distribution regression in the semi-supervised learning framework, that is where the algorithm makes use of both labeled and unlabeled data for training without posing strong restrictions.

- **Others**: Time series analysis of revenue generated from foreign tourists in India.
  Developing an algorithm for simulation of a point lying is a given polygon.
  Classification of glasses using principal component analysis and discriminant analysis.
  Analysis of loan data and finding largest possible balanced one-way layout.
  Trend filtering and its applications on grid data structure.
  Convolutional Neural Network to map images to digits in MNIST data set.

Corporate Projects

- **Project**: An analysis of economic adjustment offset for loan default data at **Upstart Network Inc.**

  Supervisors: Jessie Zaetz, Data Scientist and Don Carmichael, Manager.

  Summary: When evaluating default probability of loan borrowers, one has to take the ongoing economic situation into account. The model pipelines used to rate customers, incorporates the economic effect as an offset. The project was to fetch data from previous models used over the years, collect the offset of various pipelines, apply these offset adjustments to the present day model, and to identify factors correlated with the reflected change in the outcomes. The project required extensive application and modification of a vast private Github function repository and collaboration with various members.

- **Title**: **Analysis of loan defaulter data competition** organised by **Capital One**.

  Summary: The client was interested in identifying/predicting defaulters from a large list of loan borrowers with covariates. We used logistic regression after performing some variable transformation to identify factors important in identifying a defaulter. Also, we used survival analysis to estimate when a given loan borrower will default. This project was awarded the **first position**.

- **Project**: An analysis of **FMCG** (Fast Moving Consumer Goods) sales data provided by **Nielsen**.
  Summary: The data consisted of value offtake, number of stores, price of crude oil, price of sugar for a two year period and GDP (Gross Domestic Product) and quantities like CPI (Consumer Price Index) for a one year period. The missing values are imputed using a **SARIMA** model and polynomial regression is used to estimate the the value offtake using GDP and CPI. Also SARIMA is used again to predict the value offtake for future months.

Conference Proceedings

- Student Talk
  - Western North American Region of The International Biometric Society Conference 2021.
  - Eastern North American Region Conference 2021.
  - Joint Statistical Meetings 2020.
  - Conference on Statistical Practices 2020.
  - Joint Statistical Meetings 2019.
  - Network and Games Workshop 2016, Indian Institute of Technology, Mumbai, India.
  - Network and Games Workshop 2015, Indian Institute of Technology, Ropar, India.

- Poster Presentation
  - Pittsburgh American Statistical Association Chapter Spring 2021.
  - Innovation With Impact 2020.
  - International Conference On Statistical Distributions and Applications 2019.

Teaching (instructor or assistant)

▲ **Undergraduate level courses**

- Introduction to Statistical Inference 36-226 (Summer 2021) **(Instructor)**
- Quantitative Social Science Scholars (Spring 2021, 2022) **(Guest Instructor)**
- Modern Regression (Fall 2016, 2017, 2018)
- Introduction to Statistical Inference (Spring 2017, 2018)
- Introduction to Probability Modeling (Spring 2019)
- Advanced Data Analysis (Spring 2020)
- Engineering Statistics and Quality Control (Fall 2021)
- Probability Theory and Random Processes (Spring 2022).

▲ **Graduate level courses**

- Deep Learning (Masters of Computational Finance, Summer 2019, Fall 2019)
- Machine Learning (Masters of Computational Finance, Fall 2019).

Course Work (selected)

■ **Carnegie Mellon University** **PhD**

Intermediate Statistics	Regression Analysis
Advanced Introduction to Machine Learning	Convex Optimization
Statistical Machine Learning	Statistical Computing
Advanced Probability Overview	Foundation of Causal Inference
Advanced Statistical Theory	Teaching Statistics

■ **Indian Statistical Institute, Kolkata** **Masters**

Statistical subjects	*Mathematical subjects*
Multivariate Statistical Analysis	Measure Theory
Statistical Inference, Regression Techniques	Applied Stochastic Processes
Time Series Analysis, Operation Techniques	Metric Topology
Large Sample Statistical Methods	Image Processing

■ **Indian Statistical Institute, Kolkata** **Bachelors**

Statistical subjects	*Mathematical subjects*
Statistical Inference and Methods	Probability Theory
Linear Statistical Models, Stochastic Process	Vectors and Matrices
Design of Experiments, Numerical Analysis	Abstract Algebra
Economics (micro, macro, econometrics)	

Technical Expertise and Software(s)

- Co-author of R package **drpop** on **CRAN**. Estimates population size nonparametrically from incomplete lists with multiple options for models. Associated paper manuscript submitted.
- Programming Ability in various languages/softwares like **C++, C, R, Python (pyspark), SQL** and version control in **Git** bash.
- Experienced in Windows operating system, Microsoft Office and in GNU Linux OS.

Extra curricular activities

- Co-treasurer of Carnegie Mellon University Indian Graduate Student Association (CMU IGSA) for 2020.
- Vice-President of Carnegie Mellon University Indian Graduate Student Association (CMU IGSA) for 2018.
- General Secretary of CMU IGSA for 2017.