Shafayet Khan Shafee

sshafee@isrt.ac.bd kshafayet.netlify.app shafayetShafee shafayetshafee Google Scholar
+880-1791-051104 Dhaka, Bangladesh

Statistician and data scientist applying rigorous statistical methods to public health research and industry problems, with a commitment to reproducible research and open-source software development.

Research Interests

Causal Inference, Multilevel Modeling, Bayesian Inference, Survival Analysis, Sensitivity Analysis, Statistical Machine Learning, Public Health.

Education

M.Sc. in Applied Statistics Dhaka, Bangladesh
GPA 3.97 / 4.00 2022–2023
Institute of Statistical Research & Training
University of Dhaka

- **Thesis:** Extended the median odds ratio framework for quantifying between-cluster variation in three-level hierarchical data with a binary outcome, and developed corresponding interval estimators.
- **Selected Coursework:** Causal Inference; Multilevel Modeling; Bayesian Inference; Spatial Statistics; Statistical Machine Learning.

B.Sc. in Applied Statistics Dhaka, Bangladesh
CGPA 3.96 / 4.00 2018–2022
Institute of Statistical Research & Training
University of Dhaka

- **Project:** Examined the association between caregivers' stimulation activities and socio-emotional development in under-five children using nationally representative MICS 2019 Bangladesh data and survey-weighted logistic regression analysis.
- **Selected Coursework:** Statistical Inference; Sampling Methods; Multivariate Statistics; Generalized Linear Models; Analysis of Time Series; Design & Analysis of Experiments; Epidemiology; Survival Analysis; Econometrics; Actuarial Statistics; Industrial Statistics & Operations Research.

Professional Experience

Data Scientist Dhaka, Bangladesh
Pathao Limited Jul 2023–Present

- Applied Bayesian experimental design (Bayes Factor Design Analysis) to estimate sample sizes for controlled product experiments, incorporating prior information and evidence-based stopping criteria.
- Designed and analyzed randomized controlled experiments, including randomization checks through covariate balance assessment and Sample Ratio Mismatch (SRM) testing, to evaluate business initiatives.
- Estimated causal effects of a fintech product intervention on merchant transaction behavior using propensity score matching and Difference-in-Differences.
- Applied Hierarchical Bayesian modeling to evaluate adoption across cohorts in a controlled experiment, accounting for small and unequal group sizes and quantifying uncertainty through posterior distributions.
- Applied causal ML methods, namely Double ML and generalized random forests, to estimate heterogeneous treatment effects across user subgroups, supporting targeted intervention design.

- Co-developed a repayment risk scoring framework using Random Survival Forests to model time-to-repayment distributions, drawing on survival analysis methods for a credit risk application.
- Built scalable analytical pipelines using dbt and BigQuery, supporting reproducible statistical workflows and standardized automated analytical reporting.

Publications

† Corresponding author * Equal contribution

[1] **S. K. Shafee**†, and M. S. Rahman. *On the estimation of the median odds ratio for measuring contextual effects in multilevel binary data from complex survey designs*. [Under review]. June 2026. arXiv: 2606.15145 [stat.ME]. URL: https://arxiv.org/abs/2606.15145.

[2] **S. K. Shafee**†, B. Sarker*, and M. N. I. Sium*. *G-computation for causal effect estimation from observational hierarchical data with unmeasured cluster context*. [Under review]. June 2026. arXiv: 2606.14131 [stat.ME]. URL: https://arxiv.org/abs/2606.14131.

[3] **S. K. Shafee**†, M. N. I. Sium, B. Sarker, and R. Islam. "Investigating the causal effect of maternal continuum of care on child's minimum acceptable diet: A multilevel approach using partially pooled propensity score weighting". In: *PLOS ONE* 20.11 (Nov. 2025), pp. 1–13. DOI: 10.1371/journal.pone.0335972.

Awards & Achievements

- **Dean's Award**, *Faculty of Science, University of Dhaka* **2025**
- **Conference Award for Scientists**, for the abstract "Interval Estimation of the Median Odds Ratio for Measuring Contextual Effects in Multilevel Data Using a Binary Logistic Model", *45th Annual Conference of the International Society for Clinical Biostatistics (ISCB45), Thessaloniki, Greece* **2024**
- **National Science and Technology (NST) Fellowship**, for M.Sc. thesis research on multilevel modeling, *Ministry of Science and Technology (MoST), Government of Bangladesh* **2023**

Talks & Presentations

- "On the Estimation of the Median Odds Ratio for Measuring Contextual Effects in Multilevel Binary Data from Complex Survey Designs", *International Conference on Applied Statistics and Data Science (ICASDS 2025), University of Dhaka, Bangladesh* **2025**

Technical Skills

- **Programming:** R, Python, SQL, Lua, Octave, Julia.
- **Statistical Software:** Stata, SAS, Minitab.
- **Machine Learning:** Ensemble Models, Causal ML (DML, GRF, BCF), Model Calibration, Conformal Prediction.
- **Data & MLOps:** Shiny, BigQuery, Google Data Studio, GCP, MixPanel, dbt, Kedro, MLflow, Docker, Bash.
- **Others:** LaTeX, Git, GitHub Actions, GitLab CI.

Open Source Contributions

R Packages

- MOR — Post-estimation functions to compute the Median Odds Ratio and corresponding confidence interval from fitted multilevel binary logistic regression models.
- skmisc — Miscellaneous utility R functions.

Python Packages

- `skmiscpy` — Reusable functions for causal inference diagnostics, including mirror histograms, Love plots, and standardized mean difference (SMD) computation for inverse probability weighting (IPW).
- `kedrogen` — CLI tool for scaffolding reproducible data science projects from cookiecutter templates.

Miscellaneous

- `randomizer` — Dockerized Shiny application for random allocation of experimental units into treatment groups with specified proportions, featuring SRM testing and covariate balance diagnostics via Love plots.
- Developed multiple Quarto extensions to enhance scientific documents and presentations, including tools for code highlighting (`line-highlight`), interactive SQL demonstrations (`interactive-sql`), embedding downloadable resources (`downloadthis`), and presentation styling (`reveal-header`), available via the Quarto Extension Registry.
- `python-uv-gcloud` — Minimal Docker base image with Python, `uv`, and Google Cloud SDK for CI/CD pipelines and cloud workflows.
- `uvshot` — Bash scripts for reproducible and isolated Python environment setup using `uv`.