Haoyu Yang

haoyuyang@hsph.harvard.edu | https://haoyu-yang001.github.io/ | haoyu-yang001 | Harvard T.H. Chan School of Public Health, Biostatistics | Postdoctoral Research Fellow

Employment

Harvard University *Boston, USA*

Department of Biostatistics July. 2023 - present

Postdoctoral Research Fellow Advisor: *Xihong Lin*

Education

Renmin University of China *Beijing, China*

Ph.D. in Statistics Sep. 2018 - Jun. 2023

Tianjin University *Tianjin, China*

BS in Mathematics and Applied Mathematics; Financial Management (minor) Sep. 2014 - Jun. 2018

Research interests

Statistical Method: Large-scale Mediation Analysis; Proteomic Associations Analysis; Experimental Design; Missing and Sampling

Statistical Application: Genome-wide association study; Economics Experiment; Clinical Trial; Network Visualization

Research works

Papers:

[1] **Yang, H.**, Qin, Y., Yu, Y, Li, Y.* (2026) Design strategies for networked experiments via interference balancing. *Journal of Business & Economic Statistics*, *Accepted*, 1-31.

[2] Zhou, H., Li, Z., Shyr, D., Li, X., **Yang, H.**, Dey, R., ... & Lin, X. (2026). Comparison of variant callers using 60 532 multi-ancestry whole genome sequences. *Briefings in Bioinformatics*, 27(2), bbag130.

[3] **Yang, H.**, Wang, R.†, Lin, Y., Lin, X.* (2026+) Tail likelihood ratio method for large-scale causal mediation testing in epigenome-wide studies. *Journal of the American Statistical Association*, *Minor revision*.

[4] Fang, L., Qin, Y., Li, Y., **Yang, H*.** (2026+) Sequential Covariate-Adjusted Randomization Framework. *Statistica Sinica*, *Minor revision*.

[5] **Yang, H.**, Liu Z., Wang R., Lai E., Schwartz J., Baccarelli A., Huang Y., Lin X*. (2025) Causal mediation analysis for integrating exposure, genomic, and phenotype data. *Annual Review of Statistics and Its Application*, 12:337-360.

[6] **Yang, H.**, Qin, Y.†, Li, Y.*, Hu, F. (2024) Sequential covariate-adjusted randomization via hierarchically minimizing Mahalanobis distance and marginal imbalance. *Biometrics*, 80(2): ujae047.

[7] Qin, Y., Li, Y., Ma, W., **Yang, H.**, Hu, F.*. (2024) Adaptive randomization via Mahalanobis distance. *Statistica Sinica*, 34(1): 353-375.

[8] Li, Y., Xie, J., Qin, Y., **Yang, H.***. (2023) Network-adjusted response-adaptive randomization in economic experiment (in Chinese). *Statistical Research*, 40(10): 151-160.

[9] Li, Y., **Yang, H.**, Yu, H., Huang, H., Shen, Y.*. (2023) Penalized estimating equations for generalized linear models with multiple imputation. *Annals of Applied Statistics*, 17(3): 2345-2363.

[10] **Yang, H.**, Qin, Y.†, Wang, F., Li, Y.*, Hu, F. (2023) Balancing covariates in multi-arm trials via adaptive randomization. *Computational Statistics & Data Analysis*, 179: 107642.

[11] Li, Y., **Yang, H.**, Yu, H., Huang, H., Shen, Y.*. (2023) Penalized weighted least squares estimate for variable selection on correlated multiply-imputed data. *Journal of the Royal Statistical Society Series C*, 72(3), 703-717.

[12] **Yang, H.**, Qin, Y., Li, Y.* (2022) Group sequential randomization for split questionnaire design (in Chinese). *Systems Science and Mathematical Sciences*, 42(1):17-34.

[13] **Yang, H.**, Qin, Y., Shi, C., Ren, W., Li, Y.* (2026+) Adaptive Weighted Random Isolation (AWRI): a simple design to estimate causal effects under network interference. *Under review.*

[14] **Yang, H.**, Qin, Y., Li, Y.* (2026+) Estimation of the average treatment and spillover effects in observational studies under general interference. *Under review.*

[15] **Yang, H.**, Wang, R., Shuang, S, Lin, X.* (2026+) Synthetic Phenotype Assisted Linear Mixed Models Improve Proteome-Wide Genetic Discovery in Incomplete Biobank Data *Under Review.*

[16] **Yang, H.**, Wang, R., Lin, X.* (2026+) Heterogeneity-aware Empowerment via Adaptive Robust transfer for Target population inference (HEART-GWAS). *In progress.*

*Remark: * corresponding author; † contributed equally to the first author.*

Chapters in Books:

- Variable Selection for Missing Data. In: *Modern Survey Analysis.* Statistical Principles of Data Science. In: *Introduction to Data Science.*

Experience

Research Associate:

Research on Marketing Gas Station selection and the Marketing effect Evaluation of Oils *2022*

- Collaboration with CPPEI to develop strategies of the marketing discount for the decision-maker, and establish the evaluation process for the performance of each gas station. Based on the geographic POI information and the observational experiment data, causal inference techniques for the individual treatment effect estimation and machine learning methods are used in this project.

Responsible consumption: A study of Cognition, Behavior and Influencing factors *2020*

- Collaboration with JD to understand current Chinese consumers' cognition, behavior and attitude towards responsible consumption. Construct the responsible consumption index (RCI) to measure consumers' behavior through the questionnaire design and survey, principal component analysis, and factor analysis.

Factor Analysis of Active and Healthy Aging (Structural Analysis) *2020*

- In the project of active healthy aging, we construct the index system and further analyze the active health of the elderly based on the community-level data. The analysis methods include imputation of missing data, descriptive statistical analysis, cluster analysis, contingency analysis, correlation analysis, and inference test.

Research on Geographic Information Analysis and Sales Forecast for Gas Station *2019*

- Collaboration with CPPEI to develop an analytical model of a rating system for gas station using geographic POI information. Based on the operating data and the spatial information from gas stations to predict oil and non-oil sales using factor analysis, clustering and several classification methods.

Teaching Assistant:

2020

- Complex Statistical analysis; Introduction of Data Science

Presentations and Awards

Presentations:

2025 Statistical Genetics Meeting; 2026 ILCCO Biostat meetings; 2026 Joint Statistical Meetings *Boston, USA*

Robust Synthetic Data–Assisted Genome-Wide Association Studies Boost Proteome-Wide Genetic Discovery with Partially Observed Data in Biobanks. Nov.2025; Mar. 2026; Aug. 2026

2024 Joint Statistical Meetings - Statistics & Data Science *Portland, USA*

Tail likelihood ratio method for large-scale causal mediation hypothesis testing. Aug.2024

Statistical Genetics Meeting *Boston, USA*

Causal mediation analysis for integrating exposure, genomic, and phenotype data. Jun.2024

The 37th New England Statistics Symposium *Connecticut, USA*

Design Strategies for Networked Experiments via Interference Balancing. May.2024

The Ph.D. Forum on Biostatistical Methods and Applications *Beijing, China*

Adaptive randomization via Mahalanobis distance. Jun.2021

Awards:

Jun. 2026	Poster Award, 2026 ICSA Applied Statistics Symposium
Oct. 2022	Top Ten Papers of the 8-rd National Postgraduate Statistics Forum
Aug. 2021	Top Ten Papers of the 2-rd Data Science and Modern Economic Statistics Forum
Dec. 2020	First Price of "BeiGene" Excellent Paper for Youth

Teaching:

2021	Teaching Assistant — Complex Statistical Analysis, Probability Theory
2020	Teaching Assistant — Introduction to Data Science

Journal Review

- Statistica Sinica, Biostatistics, Journal of Data Science