Out of the Matrix: Utilizing Data Simulations for Public Health Research

S. J. Robbins[1], K. Roper[2], N. Kale[2], A. Mangino[1]

[1]Department of Biostatistics, [2]Department of Family Medicine

College of Public Health, University of Kentucky, Lexington, KY

BACKGROUND

Simulations mirror real-world scenarios or conditions from the patterns of data.[1]

Simulation is a computation tool for many tasks, such as examining random variables, independence, discrete and continuous distributions, confidence intervals, hypothesis testing, and efficient estimators. [1,2]

Common statistical softwares that are utilized across public health systems, such as R and SAS, have the capability of performing data simulations. [3,4]

Despite the benefit of simulation, it is often underutilized by public health researchers. [5,6,7]

In this motivating example, practical applications of simulations for public health analyses are illustrated using simulated data by biostatisticians in the Biostat CIRCL. The goal is to demonstrate the ease and convenience of data simulations for public health research.

CASE STUDY

The goal of this project was to simulate the expected responses from the Society of Teachers of Family Medicine's (STFM) CAFM Educational Research Alliance (CERA) annual survey for 2023. [8]

The motivation behind this simulation was to expedite coding and analyses due to 90 days of research exclusivity before data became publicly available.

General membership questionnaires were provided and included demographic variables such as age, race, gender, and regional location.

The survey module (developed by Biostat CIRCL) included categorical response options, and included questions pertaining to trust, communication, practice and education after the Dobbs vs. Jackson ruling.

Continuous variables were simulated randomly from a normal distribution with the means and standard deviations from previous year estimates. Categorical variables were randomly simulated based on the number of categories and the estimate proportions in each group. Estimated proportions for the new survey module are assumed to be random.

Comparisons between the simulated data and the “real” data are seen in the results section. Information on the actual data is limited due to forthcoming publication.

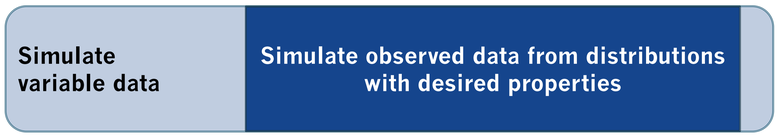
Simulations were performed in R version 4.2.1.

THE SIMULATION

The following figure aims to demonstrate the workflow in creating a simulated dataset from known and unknown estimates. Estimates were identified from literature from three sources. [9,10,11]

For this case study, previous literature suggested that the average response rate was 20% for the 5000 potential responses. 1000 subjects were selected.

Categorical and continuous estimates were derived from demographic table results. Unknown variables represent the 10 new questions for this project.



Observations were drawn from previous proportion and means from demographic data. New variables were drawn at random.

RESULTS

Variable	Simulated Data N = 1000	2023 CERA Data N = 1198
Age		
	44.35 (10.84)	**48.2** (12.2)
Gender		
Male	**395** (39.5%)	**427** (35.6%)
Female	**605** (60.5%)	**750** (62.6%)
Race		
White	**832** (83.2%)	**902** (75.3%)
Black or African-American	**32** (3.2%)	**57** (4.8%)
Asian	**89** (8.9%)	**111** (9.3%)
Region		
1	**58** (5.8%)	**76** (6.3%)
2	**132** (13.2%)	**134** (11.2%)
3	**186** (18.6%)	**204** (17.0%)
4	**33** (3.3%)	**42** (3.5%)
5	**180** (18.0%)	**229** (19.1%)
6	**61** (6.1%)	**86** (7.2%)
7	**91** (9.1%)	**139** (11.6%)
8	**102** (10.2%)	**106** (8.8%)
9	**157** (15.7%)	**182** (15.2%)

Table 1. Results support simulation data’s potential to mimic real-world data when variables are simulated from parameter estimates.

CONCLUSION

Data can be simulated to reflect previous, known findings to build analytical frameworks for efficient research workflows.

With the growing accessibility of statistical software in public health systems, there are potential instances where data simulations could be utilized as a tool for public health data teams.

Simulated data presents options for public health practitioners to have a practical method of analysis prepping and/or auditing for annual, protected or large datasets.

Incorporating more advanced methods of simulation in public health are possible, including:

- *primary analysis comparison methods,*
- *pilot testing counterfactual outcomes,*
- *justification of selected methodologies for publications*

ACKNOWLEDGEMENTS

Interested in simulations in R? Want to see the full code? Use your phone on the QR code to download a PDF with the code used for this simulation.

The Biostatistics Consulting and Interdisciplinary Research Collaboration Lab (Biostat CIRCL) is a group of biostatisticians who specialize in balancing rigorous statistical methodology with the complex challenges of interdisciplinary biomedical research. Our goal is to work with researchers to build effective research teams in order to optimize data-driven discoveries.
I want to thank Drs. Emily Slade, Tony Mangino and Amanda Ellis for their support on this project.

REFERENCES

1. Jennings, W. (2022, July 13). *Data simulation: Tools, benefits, and use cases*. Gretel. Retrieved March 30, 2023, from https://gretel.ai/blog/data-simulation
2. Robb, D. (2021, May 21). *What is data simulation?* Datamation. Retrieved March 30, 2023 from https://www.datamation.com/big-data/data-simulation/
3. R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
4. SAS Institute Inc 2013. SAS/ACCESS® 9.4 Interface to ADABAS: Reference. Cary, NC: SAS Institute Inc
5. Mantica, G. (2021, December 10). *How can simulation modeling inform public health?* Boston University. Retrieved March 30, 2023 from https://www.bu.edu/hic/2021/12/10/how-can-simulation-modeling-inform-public-health/
6. Boulesteix A., Groenwold R., Abrahamowicz M. (2020). Introductions to statistical simulations in health research. *BMJ Open. doi: 10.1136/bmjopen-2020-039921*
7. Maglio, P.,Sepulveda, M., Mabry, P. (2014). Mainstreaming modeling and simulation to accelerate public health innovation. *American Journal of Public Health.* 104. 1181-1186. https://ajph.aphapublications.org/doi/abs/10.2105/AJPH.2014.301873
8. About CERA (2023). Accessed March 30, 2023. https://www.stfm.org/publicationsresearch/cera/cera/
9. Thai JN, Saghir HA, Pokhrel P, Post RE. Perceptions and Experiences of Family Physicians Regarding Firearm Safety Counseling. Fam Med. 2021;53(3):181-188. https://doi.org/10.22454/FamMed.2021.813476.
10. Krys E. Foster, Allison R. Casola, Zeynep Uzumcu, Sascha Wodoslawsky & Christina Kelly (2022) Outpatient maternity care and telemedicine use perceptions in the COVID-19 pandemic: a 2020 CERA survey, Women & Health, 62:5, 402-411, DOI: 10.1080/03630242.2022.2072051
11. Dichter ME, Teitelman A, Klusaritz H, Maurer DM, Cronholm PF, Doubeni CA. Trauma-Informed Care Training in Family Medicine Residency Programs: Results From a CERA Survey. Fam Med. 2018;50(8):617-622. https://doi.org/10.22454/FamMed.2018.505481.