Abstract. [14] discover a non-Gaussian asymptotic distribution in rerandomization theory of possible relevance in philosophy of statistics.

Key words and phrases: asymptotic balance criteria, robustness-balance tradeoff, Mahalanobis distance.

1. INTERPRETATION

Fisher suggests "the question 'Of what population is this a random sample?' ... must frequently be asked by every practical statistician" [7, Section 2]. Randomized experiments for causal effects should balance confounders between members in treatment and control groups for principled [18, 20] inference of causal effects. The simplest observed dataset for a randomized experiment (without principal stratification [8]) collects actual binary assignment and outcomes usually for an average treatment effect estimand. The actual binary assignment can be interpreted as the sole contribution to actual relevant causal presence (absence), thereby allowing for an inferential emphasis on randomization [22, 11, 2]. The question arises of which population is the actual binary assignment vector a random sample? Such an assignment population would clarify aspects of actual causal connection in the experiment over the N members, providing improved understanding of causal aspects of the experiment. The assignment population is a set of possible assignments with cardinality $2^N - |R|$ where $|R|$ is the cardinality of assignments which fail the covariate balance criteria.

Fisher suggests the possibility of general causal understanding in statistical science: "any set of independent measurements ... from an infinite population ... are a random sample from the totality of numbers produced by the same matrix of causal conditions" [7, Section 2]]. Rerandomization is "widely used in practice" [19, Section 1] since it provides a unified framework for blocking, matching, and restricted randomization [19, Section 5.1]. By discarding imbalanced assignment vectors using pre-specified balance criteria [19, Section 1] rerandomization can clarify the assignment population cardinality and distribution, and therefore rerandomization can clarify aspects of the "causal conditions" expressed by the actual binary assignment of causal connections. If N is finite then balance on key pre-treatment covariates can exhibit causal understanding, but robustness issues complicate this causal interpretation since requiring balance over multiple key covariates could lead to large $|R|$. If N approaches infinity then robustness issues seem less central since 2^N also approaches infinity. [10] provide a method to optimize the robustness-balance tradeoff in randomized experiments and argue that the tradeoff can vanish in large samples. Then, as 2^N is infinite, the notion of discarding imbalanced assignments through rerandomization balance criteria can be considered more directly related to general causal understanding in randomized experiments through the question "of what population is this a random sample?" Perhaps randomized experiments are a more suitable basis than observational studies from which to develop philosophical insight about general statistical causality [23, 26, 1, 21].

[14] use the Mahalanobis distance to uncover a non-Gaussian asymptotic distribution in rerandomization theory: "a linear combination of a Gaussian random variable and truncated Gaussian variables". [24] describe the discovery as "a quite complex asymptotic sampling distribution". The Mahalanobis distance is a useful balance criteria "because it is an affinely invariant scalar measure of multivariate covariate balance" [19, Section 3]. [12, Section 2.3.3] critiques the rerandomization approach of [19] as "common, but arguably historically haphazard". [11] critique the optimality approach of [12] for emphasizing member-level variation instead of assignment-level variation in randomization-based inference. [11] defend the use of Mahalanobis-based balance criteria for rerandomization from [12]' critique noting that "non-linear dependencies also can be considered in Mahalanobis-based rerandomization" (p. 400). [5] suggest a modified Mahalanobis distance for settings with collinearities or high-dimensional covariates but state "the theory developed in [14] cannot be readily applied to ridge rerandomized experiments" (p. 296). [28] suggest a modified Mahalanobis approach using principal component analysis for improved high-dimensional covariate balance but do not study "the asymptotic property of the treatment effect estimator under PCA rerandomization" (p. 21). [24] use a relation between high-reject balance criteria in asymptotic settings and optimality considerations from [12] to argue that "the asymptotic sampling distribution under Mahalanobis-based rerandomization simplifies to a normal distribution" for optimal designs; this result emphasizes the importance of the notion of asymptotic acceptance probability for rerandomization theory [14, p.9158].

The asymptotic sampling distribution of [14] seems interpretable as a discovery about the 'shape' [9] of causal connection in randomized experiments (without principal stratification) and perhaps generally about the 'shape' of "causal conditions" [7, Section 2] in statistical science. [14] show rerandomization asymptotically outper-

forms randomization by lowering the "asymptotic quantile ranges of the difference-in-means estimator" which could imply that rerandomization procedures can exhibit general causal understanding. Covariate balancing criteria can provide some relevant causal understanding [3] since in randomized experiments they determine the relevant assignment population from which the actual assignment of causal connection is drawn. Randomization over assignment population is the core inferential tool in randomized experiments for causal effects, so rerandomization-based considerations which impact the cardinality of assignment populations must be essentially causal considerations since the causal science is supposed to be fully in the randomization over assignment population [22]. Covariate balance criteria determine the interpretation of the actual assignment vector and the interpretation of the randomization-based inference. This causal interpretation of covariate balance criteria in randomized experiments is complicated by tradeoffs with robustness considerations, but these robustness complications are simplified in large samples [10] allowing for valid general causal interpretation of asymptotic covariate balance criteria. The asymptotic sampling distribution of [14] may provide useful basis for inquiry about the role of 'actual causal connection' in statistical causal inference and statistical science.

Perhaps basic inquiry about the asymptotic sampling distribution of [14] can help raise further interesting methodological questions in the developing asymptotic rerandomization theory. [16] study asymptotic aspects of rerandomization when assignment and balance require cluster-level considerations. [15] study asymptotic aspects of rerandomization with regression-adjustment. [4] study asymptotic statistical power of tests in rerandomization. [27] study asymptotic rerandomization with increasingly strict balance criteria and increasing number of covariates. General statistical causality from asymptotic rerandomization may relate broadly to basic aspects of problems of specification [17, 13] since problems of specification are closely related to the "causal conditions" of well-specified populations [7, 6, 25].

REFERENCES

[1] BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–973.

[2] BASSE, G. W., FELLER, A. and TOULIS, P. (2019). Randomization tests of causal effects under interference. *Biometrika* **106** 487–494.

[3] BEN-MICHAEL, E., FELLER, A., HIRSHBERG, D. A. and ZUBIZARRETA, J. R. (2021). The Balancing Act in Causal Inference.

[4] BRANSON, Z., LI, X. and DING, P. (2024). Power and sample size calculations for rerandomization. *Biometrika* **111** 355–363.

[5] BRANSON, Z. and SHAO, S. (2021). Ridge rerandomization: An experimental design strategy in the presence of covariate collinearity. *J. Stat. Plan. Inference* **211** 287–314.

[6] D'AMOUR, A., DING, P., FELLER, A., LEI, L. and SEKHON, J. (2021). Overlap in observational studies with high-dimensional covariates. *J. Econom.* **221** 644-654.

[7] FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. A* **222** 309-368.

[8] FRANGKAKIS, C. E. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29.

[9] GÖRTZ, U. and WEDHORN, T. (2010). *Algebraic Geometry I: Schemes*. Springer.

[10] HARSHAW, C., SÄVJE, F., SPIELMAN, D. A. and ZHANG, P. (2024). Balancing Covariates in Randomized Experiments with the Gram–Schmidt Walk Design. *J. Am. Stat. Assoc.*

[11] JOHANSSON, P., RUBIN, D. B. and SCHULTZBERG, M. (2021). On Optimal Rerandomization Designs. *J. R. Stat. Soc. Series B* **83** 395–403.

[12] KALLUS, N. (2018). Optimal a priori balance in the design of controlled experiments. *J. R. Stat. Soc. Series B* **50** 3439–3465.

[13] LEHMANN, E. L. (1990). Model Specification: The Views of Fisher and Neyman, and Later Developments. *Stat. Sci.* **5** 160–168.

[14] LI, X., DING, P. and RUBIN, D. B. (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proc. Natl. Acad. Sci.* **115** 9157–9162.

[15] LI, X. and DING, P. (2020). Rerandomization and Regression Adjustment. *J. R. Stat. Soc. Series B* **82** 241–268.

[16] LU, X., LIU, T., LIU, H. and DING, P. (2022). Design-based theory for cluster rerandomization. *Biometrika* **110** 467–483.

[17] MCCULLAGH, P. (2002). What is a statistical model? *Ann. Stat.* **30** 1225–1310.

[18] MENG, X.-L. (2018). Conducting highly principled data science: A statistician's job and joy. *Stat. Probab. Lett.* **136** 51–57.

[19] MORGAN, K. L. and RUBIN, D. B. (2012). Rerandomization to improve covariate balance in experiments. *Ann. Stat.* **40** 1263–1282.

[20] OGBURN, E. L. and SHPITSER, I. (2021). Causal Modelling: The Two Cultures. *Obs. Stud.* **7** 179–183.

[21] ROBINS, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Model.* **7** 1393-1512.

[22] RUBIN, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *Ann. Stat.* **6** 34–58.

[23] RUBIN, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat. Med.* **2** 20–36.

[24] SCHULTZBERG, M. and JOHANSSON, P. (2020). Asymptotic Inference for Optimal Rerandomization Designs. *Open Stat.* **1** 49–58.

[25] SUR, P. and CANDÈS, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *J. Econom.* **116** 14516–14525.

[26] VANDERWEELE, T. J. (2019). Principles of confounder selection. *Eur. J. Epidemiol* **34** 211–219.

[27] WANG, Y. and LI, X. (2022). Rerandomization with Diminishing Covariate Imbalance and Diverging Number of Covariates. *Ann. Stat.* **50** 3439–3465.

[28] ZHANG, H., YIN, G. and RUBIN, D. B. (2024). PCA Rerandomization. *Can. J. Stat.* **52** 5–25.