STATISTICS · CAUSAL INFERENCE

Randomised Controlled Trials as the Benchmark for Causality

Why random assignment gives the most credible answer to cause-and-effect questions, what a well-run trial needs, and where randomised experiments reach their limits in business and survey research.

ModQu research article · 1,031 words · 10 references (APA 7th edition)

When the question is "does this work?", a **randomised controlled trial** (RCT) is the standard against which other evidence is judged. Participants, customers or sites are assigned to a treatment or a control condition by chance, and outcomes are compared. The idea is simple, but it solves the central problem of causal inference more completely than any other design.

Why chance assignment works

A causal effect compares what happened to a group with what would have happened to the same group without the treatment (Holland, 1986). Because that counterfactual can never be observed, every causal study needs a comparison group that stands in for it. In observational data, the people who receive a treatment usually differ from those who do not: customers who opt in to a premium service are more engaged, and stores chosen for a pilot are often the best managed. These differences, or confounders, are mixed up with the treatment effect.

Random assignment breaks the link. Because a coin, not a person's characteristics, decides who is treated, the groups differ only by chance in everything that existed before the trial, whether measured, unmeasured or unknown. Any systematic difference in outcomes can then be attributed to the treatment. Fisher (1935) also showed that randomisation provides the basis for the statistical test: the observed difference can be compared with the differences that would arise from all the other random assignments that might have occurred.

Chance imbalances still happen, but they shrink as the trial grows, and confidence intervals account for them. Figure 1 compares randomised groups with self-selected ones. With 50 customers per group, the share of loyal customers could differ by up to about 19 points purely by chance. With 5,000 per group, the range is under 2 points. A self-selected comparison with a 15-point gap keeps that gap however much data are collected.

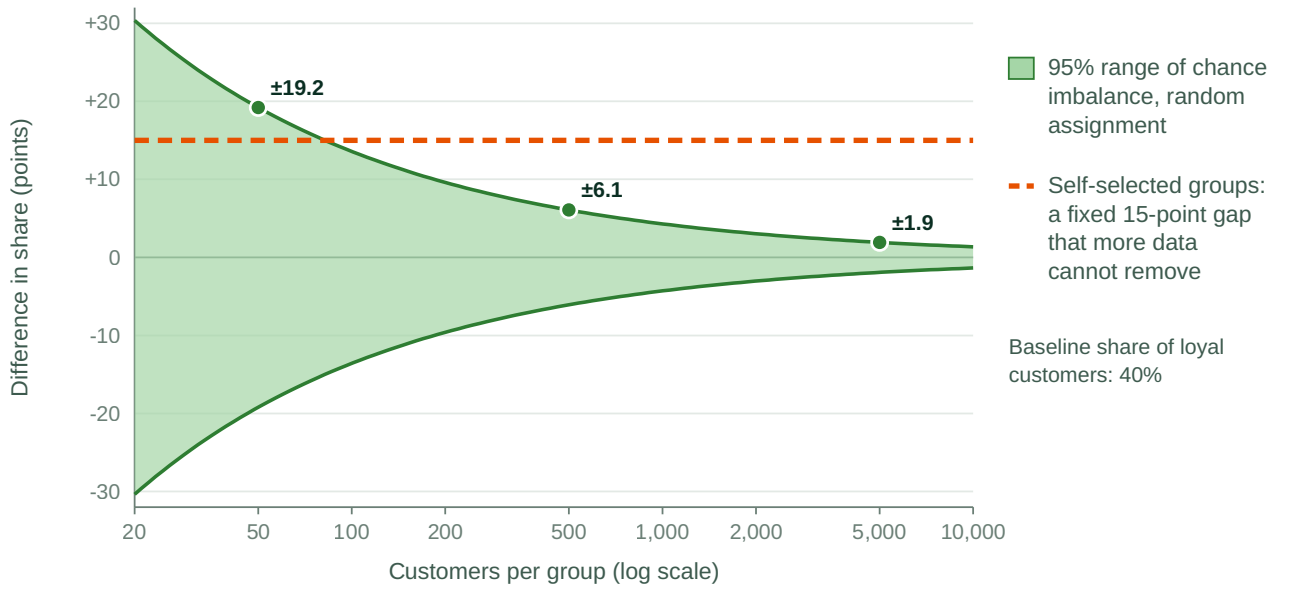


Figure 1. Under random assignment, imbalance in a baseline trait (here a 40% share of loyal customers) is due to chance and narrows as samples grow. Under self-selection, a systematic gap of 15 points persists. Illustrative data; chart by ModQu.

Evidence that the benchmark matters

The Medical Research Council's (1948) trial of streptomycin for pulmonary tuberculosis is often cited as the first properly randomised clinical trial, with allocation by random numbers held in sealed envelopes. Since then, comparisons between randomised and non-randomised estimates of the same effect have repeatedly shown why the benchmark matters. LaLonde (1986) took the results of a randomised job-training experiment and tried to reproduce them with standard econometric methods using non-experimental comparison groups. Many of the non-experimental estimates were far from the experimental result, and some had the wrong sign. In marketing, Gordon et al. (2019) found that observational methods applied to large advertising experiments at Facebook often misestimated the true effect of the ads, usually overstating it.

What a good trial needs

Randomisation guarantees comparable groups only at the moment of assignment. Several practices protect that comparability afterwards.

- **Concealed allocation.** Those enrolling participants must not know or influence the next assignment. Schulz et al. (1995) found that trials with inadequately concealed allocation reported larger treatment effects than adequately concealed trials, by about 40% on average on the odds ratio scale.

- **Blinding where possible.** Participants, staff and outcome assessors who know the assignment may behave or judge differently.
- **Intention-to-treat analysis.** Units are analysed in the group to which they were randomised, even if they did not comply. Dropping non-compliers reintroduces self-selection.
- **A pre-specified primary outcome and sample size.** Deciding in advance what will be measured, and how many units are needed, prevents searching for a favourable result.
- **Transparent reporting.** The CONSORT statement sets out the items a trial report should include, from the randomisation method to the flow of participants through the study (Schulz et al., 2010).

Randomised experiments in business

Online A/B tests are RCTs: visitors are randomly assigned to versions of a page, price or message, and the difference in conversion is the causal effect of the change. Large firms run thousands of these experiments each year, and the practical lessons, including sample ratio checks, guardrail metrics and the danger of stopping early, are summarised by Kohavi et al. (2020). In a hypothetical test with 20,000 visitors per version, a conversion rate of 4.4% against 4.0% gives an estimated lift of 0.4 percentage points, with a 95% confidence interval of roughly 0.0 to 0.8 points. Because assignment was random, the interval reflects chance variation only, not systematic differences in who saw each version. Survey researchers use the same logic when they randomly assign respondents to different question wordings, product concepts or price points within a questionnaire.

Limits of the benchmark

An RCT is the benchmark for internal validity, not a guarantee of useful answers.

- **External validity.** A trial estimates the effect for the people, setting and period studied. Rothwell (2005) shows how selective eligibility criteria and settings can limit how far trial results apply to routine practice. A test run in one market or season may not transfer to another.
- **Average effects hide variation.** The trial estimates an average. It does not show who benefits, and subgroup findings need replication.
- **Feasibility and ethics.** Some treatments cannot be randomised, such as a company-wide rebrand, and some should not be.
- **Interference.** When treated units affect controls, as in referral schemes or shared marketplaces, the simple comparison is biased.
- **Not the only source of evidence.** Deaton and Cartwright (2018) argue that RCTs do not automatically outrank other evidence. Their results need theory and background knowledge to be interpreted and applied elsewhere.

Key takeaways

Random assignment is the most reliable way to create comparable groups, because it balances unmeasured as well as measured factors and makes any remaining imbalance a matter of quantifiable chance. That is why RCTs serve as the benchmark against which observational methods are tested. Their value depends on sound execution, including concealed allocation, intention-to-treat analysis and pre-specified outcomes, and their results should be applied with attention to who was studied, where and when.

References

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine, 210*, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

Fisher, R. A. (1935). *The design of experiments*. Oliver and Boyd.

Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science, 38*(2), 193–225. https://doi.org/10.1287/mksc.2018.1135

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354

Kohavi, R., Tang, D., & Xu, Y. (2020). *Trustworthy online controlled experiments: A practical guide to A/B testing*. Cambridge University Press. https://doi.org/10.1017/9781108653985

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review, 76*(4), 604–620.

Medical Research Council. (1948). Streptomycin treatment of pulmonary tuberculosis: A Medical Research Council investigation. *British Medical Journal, 2*(4582), 769–782. https://doi.org/10.1136/bmj.2.4582.769

Rothwell, P. M. (2005). External validity of randomised controlled trials: "To whom do the results of this trial apply?" *The Lancet, 365*(9453), 82–93. https://doi.org/10.1016/S0140-6736(04)17670-8

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMJ, 340*, Article c332. https://doi.org/10.1136/bmj.c332

Schulz, K. F., Chalmers, I., Hayes, R. J., & Altman, D. G. (1995). Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA, 273*(5), 408–412. https://doi.org/10.1001/jama.1995.03520290060030