The Gold Standard of Causal Inference: Experimental Methods

Rui Huang* and Pian Chen** | June 12, 2026

I. Introduction

This article is the first part of our data science crash courses — *Causal Methods in the Courtroom and the C-Suite*. The courses introduce fundamental methods that economists, statisticians, and data scientists use to make causal inference. We write for two groups: (1) lawyers who want to understand whether challenged conduct caused harm and how to quantify damages and (2) business leaders who make product, pricing, and marketing decisions. This article focuses on experimental methods for causal inference. Later installments in this series will cover the statistical, econometric, and machine-learning methods that have been tested in the courtroom or have been used by business leaders.

Nearly every consequential decision in the courtroom and in business hinges on what-if questions: what would have happened in a world where the thing we're evaluating didn't occur? In antitrust, consumer protection, and other litigation matters, experts need to analyze alleged conduct (e.g., price fixing, consumer fraud, IP infringement) and quantify the difference between what actually happened and what would have happened absent the alleged conduct. In business decisions, millions or billions of dollars could be at stake if a different product is offered, prices are set differently, and advertising channels and spending change. Throughout this series, we use marketing and advertising to illustrate the methods.

II. Marketing and Advertising Spend

Every quarter, in nearly every marketing organization, the CFO asks some version of the same question: if we cut this advertising channel by 20%, what happens to revenue? In most cases, the honest answer is that the company does not actually know. The dashboards report only what the platforms themselves report: Meta says it drove $4M last quarter; Google says $6M. The attribution tool combines these figures and assigns a single, tidy number to each channel.

The difficulty is that those numbers add up to more revenue than the company actually earned. Even after they are reconciled, they answer the wrong question. They identify which channel touched a customer who eventually purchased; they do not reveal whether that customer would have purchased anyway. This is the distinction between attributed revenue (which channel received credit) and incremental revenue (which channel actually caused the purchase). Cutting a channel that looks strong on the dashboard but is mostly claiming credit for organic demand has little impact on the company's revenue. But cutting a channel that looks mediocre but is in fact acquiring new customers can be a serious and expensive mistake.

The same question shows up everywhere a leader must decide whether something is working. Did the price cut reduce churn, or did it just coincide with a quiet quarter for customer switching? Did the training program improve staff productivity, or did we send the best performers to the training first? Every one of these is the same problem: distinguishing what the intervention caused from what would have happened anyway. The dashboards always show that something

happened. But the causal question is whether the intervention is what made it happen. How can a decision-maker know with confidence? Where it is feasible, the cleanest answer is to run a controlled experiment.

III. Why Experiments Are the Gold Standard

To assess the effect of challenged conduct, we must compare the world in which that conduct occurred with a world identical *in every respect* except that the conduct didn't happen. The first world is what we observe. The second world is the counterfactual, which we do not observe, and it's where all the difficulty lives. The entire field of causal inference is built around methods for constructing a credible estimate of that unobserved counterfactual, and the methods differ mainly in underlying assumptions.

Constructing the counterfactual is challenging because of two fundamental problems:

- **Problem 1**: Too many other things are moving at the same time. Sales rose 15% after the campaign launched. Maybe the campaign did it — or maybe March is the strong season, or prices were cut, or a competitor stumbled. A naïve before-and-after comparison credits the campaign for movement that had other causes.
- **Problem 2**: The exposed and unexposed groups may not be comparable. The people who saw the ad differ from people who didn't, and the ad platform deliberately shows the ad to the people most likely to buy. Comparing exposed to unexposed mostly measures who the algorithm selected, not what the ad did.

A randomized experiment addresses both problems at once — random assignment. Because users are assigned to treatment (i.e., exposed to the ads campaign) or control group (i.e., not exposed to the ads campaign) by a coin flip, the two groups are statistically identical before the ad runs — same income mix, same purchase intent, same everything, observed or not. This eliminates Problem 1 because the difference between the treatment and control is not caused by seasonality, competitive moves, or anything else changing over time. Whatever those forces do, they hit both groups equally and cancel out in the comparison. It eliminates Problem 2 as well because users do not choose their group, and the algorithm does not hand-pick the exposed.

This is what makes randomized experiments so powerful, and so much easier to defend than any observational method: there is no need to model the confounders or to match treated and control units on their observable characteristics, and no assumptions are required about how customers behave. Randomization makes both problems disappear by construction.

The same logic is why courts, regulators, and the FDA treat randomized trials as the highest tier of evidence. A litigation expert defending a damages model built on observational data must defend the assumptions that prevent Problems 1 and 2 from corrupting the answer. An expert whose conclusions rest on a randomized experiment doesn't carry that burden — the design carried it for them. Every other method in later parts of this data science course is, at heart, a way of trying to recover what an experiment would have shown, when running the experiment itself isn't possible.

IV. Three Randomized Experiments for Measuring Advertising Effects

A few forms of randomized testing dominate the modern measurement of advertising effects. They differ chiefly in the practical constraint each is designed to overcome — in particular,

whether the advertiser can randomize at the level of individual users or only across larger units such as geographic markets.

1) A/B Test

An A/B test randomly sorts people into two groups that are identical in every way except one: only the treatment group gets the intervention being tested, while the control group does not. Platform-managed A/B tests of this kind are offered by Meta, Google, and most large advertising platforms. Before a campaign runs, the platform randomly assigns eligible users to two groups: treatment users see the ad as normal, while control users would have qualified to see it but are withheld. After several weeks, the platform compares conversion rates in the two groups. The conversion rate is the share of users who take the desired action — a purchase, sign-up, or download. The conversion lift is the difference in conversion rates between the treatment and control groups. It captures the incremental effect of advertising and is the cleanest experimental measure available to a single advertiser.

2) Ghost Ad

Early lift studies suffered from a subtler version of Problem 2. Platforms compared users who saw an ad against a control group that included many users who would never have been targeted in the first place. Because those users were never candidates to see the ad, including them in the control group diluted the measured effect and understated the campaign's true lift.

Johnson, Lewis, and Nubbemeyer at Google introduced the ghost ad to solve the problem in 2017.[1] Under this approach, the platform runs its full targeting and auction process for the control group exactly as it does for the treatment group, identifies the specific users who would have seen the ad, and records a "ghost" impression for each of them without actually serving it. What does a ghost impression mean? Suppose a retailer runs a campaign for running shoes. When a control-group user opens an app, the platform runs its usual targeting and auction and determines that this user would have been shown the shoe ad. Rather than serving it, the platform shows whatever the user would otherwise see and silently logs a "ghost impression" — a record that says "this user would have seen the ad here." Weeks later, purchases among treatment users who actually saw the ad are compared with those among control users who logged a ghost impression.

The measured effect is then the difference between the users who saw the ad and the matched users who would have seen it but for their assignment to the control group. Because both groups cleared the identical targeting and auction bar, they are nearly indistinguishable except for the ad itself, so the gap in their purchase rates is the campaign's true lift. This technique underlies the lift studies run by every major platform today.

3) Geo experiments

Not every important channel can be split at the user level. Connected TV is improving but still partly broadcast. Out-of-home billboards reach whoever drives by. Linear TV reaches whole households. And privacy changes (e.g., Apple's App Tracking Transparency, third-party cookie deprecation) have made user-level measurement steadily harder even in pure digital channels for any advertiser who isn't Google or Meta.

The fix is to do geo experiments, which randomize places instead of people.[2] The campaign is turned on in one set of designated market areas (DMAs) such as Phoenix, Charlotte, Indianapolis and turned off in a matched set. After a few weeks, sales in the treated markets are compared with those in the control markets. The comparison is clean for the same reason A/B tests are clean: the assignment is random, so the cities are statistically comparable except for the campaign.

An influential early example is a 2015 Econometrica paper by Blake, Nosko, and Tadelis, which found that branded paid search generated little incremental value for an established advertiser like eBay. The researchers measured conversion lift by having eBay switch off its branded paid search across roughly a third of U.S. markets — a geo experiment in all but name.[3] The estimate did not come from a regression coefficient; it was the difference between treated and control DMAs after randomization.

Geo experiments have become the workhorse of modern marketing measurement. They work for any advertising channel — TV, CTV, podcasts, direct mail, regional digital. Geo experiments work even inside walled gardens, where individual user data is not accessible to the advertiser — they require only a signal of whether the campaign ran in each market. They measure total business impact, including word-of-mouth and offline conversion, and they're privacy-safe. The trade-offs: there are only ~210 DMAs in the U.S., so sample size is limited; treated and control cities must be reasonably comparable before the test; and spillovers (a Charlotte resident who drives to Atlanta) are a constant concern.

Meta open-sourced GeoLift to handle much of the design and analysis.[4][5] Google has an analogous toolkit inside Google Ads.[6] Third-party vendors like Haus, Measured, and Recast will run geo tests for advertisers who don't want to build the capability in-house.[7]

V. Conclusion and What's Coming Up Next

Where experiments are feasible, they're the best available tool, and they should be the foundation of any serious measurement program — for marketing spend, for product launches, for pricing changes, for operational interventions, for anything where business leaders have to decide whether the thing they did is what made the number move. In the marketing context specifically, for most companies in 2026 the right starting point is: A/B test the digital channels where the platforms support it (today, these platform lift tests are typically ghost-ad designs); geo-test everything else. That covers something like 90% of the typical marketing budget with experimental rigor.

But experiments can't answer every question. They can't reveal the shape of the response curve, what would happen at half the budget, how channels interact, or what a campaign did three years ago when nobody ran a test. They also can't be retrofitted to a one-time event that's already happened: a market launch, a competitor's exit, a regulatory change, or a merger whose effects are now baked into market prices. Those questions require the observational toolkit.

In later parts of the data science courses, we will provide a high-level overview of the observational toolkit, including difference-in-differences, synthetic control, propensity score matching, regression discontinuity, double machine learning, and the survey-based methods that handle questions about things that don't yet exist. Each of these methods is, in its own way, an effort to approximate what a randomized experiment would have shown. Each earns its credibility by how convincingly it addresses the two problems described above — confounding factors and non-comparable groups. And each is ultimately measured, in the courtroom and the boardroom alike, against the experiment that could not be run.

Sources

[1] Johnson, G. A., Lewis, R. A., & Nubbemeyer, E. I. (2017). "Ghost Ads: Improving the Economics of Measuring Online Ad Effectiveness." *Journal of Marketing Research*, 54(6): 867-884. https://doi.org/10.1509/jmr.15.0297

[2] Vaver, J., & Koehler, J. (2011). "Measuring Ad Effectiveness Using Geo Experiments." *Google Research*. https://research.google/pubs/measuring-ad-effectiveness-using-geo-experiments/

[3] Blake, T., Nosko, C., & Tadelis, S. (2015). "Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment." *Econometrica*, 83(1): 155-174. https://doi.org/10.3982/ECTA12423

[4] 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

[5] Meta (Facebook Incubator). GeoLift: An Open-Source Tool for Geo-Based Incrementality Testing. https://github.com/facebookincubator/GeoLift

[6] Google Ads Help. Conversion Lift Based on Geography. https://support.google.com/google-ads/topic/16104796?hl=en&ref_topic=16104405&sjid=6298107622762726685-NA

[7] See, e.g., Haus (haus.io), Measured (measured.com), and Recast (getrecast.com), third-party providers of geo-based incrementality experiments.

About the Authors

* **Dr. Rui Huang** is Principal Economist and Data Science Expert at Nutcracker Economics. She has over 20 years of experience in antitrust economics, applied econometrics, and data science, with expertise spanning causal inference, experimentation design, and marketing ROI measurement. She spent over eight years leading science teams at major technology companies, served as a Staff Economist at the U.S. Department of Justice Antitrust Division, and was a tenure-track professor at the University of Connecticut with 10+ peer-reviewed publications. She holds a PhD in Economics from UC Berkeley.

** **Dr. Pian Chen** is Founder and Lead Economist at Nutcracker Economics. She has over 15 years of experience in litigation consulting and government oversight, with deep expertise in antitrust economics and financial fraud investigation. She previously held senior positions at leading litigation consulting firms and served as Associate Director of Economic Modeling at the Public Company Accounting Oversight Board (PCAOB). She holds a PhD in Agricultural and Resource Economics from UC Davis.