CAUSAL INFERENCE · EXPERIMENTAL DESIGN · CUSTOMER ANALYTICS

Did the Campaign Work? *reading cause from a randomized experiment*.

Estimating the causal effect of a marketing treatment from a 64,000-customer randomized experiment: the lift with a confidence interval, the balance check that says the lift is causal, who responds most, and the design math that sizes the test.

Author S. Ize-Iyamu **Audience** Marketing & customer analytics **Length** 2 pages **Method** RCT · inference · power

Stack Python · statsmodels · scipy

The Problem

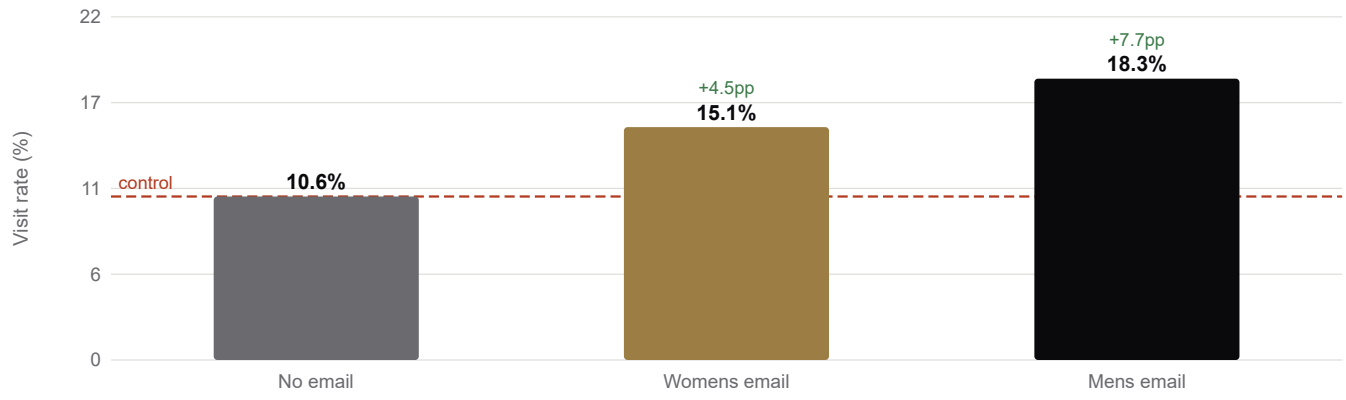
Customers who got an email visited more, but that alone proves nothing. The people you choose to email are usually your best customers, so a raw before-and-after comparison confounds the campaign with who was targeted. A randomized experiment breaks that link: assign the treatment by coin flip and the treated and untreated groups are the same on everything except the email, so the difference in outcomes is the causal effect, not a selection artifact.[1] This brief works a public 64k-customer email experiment[2] end to end, from the effect to the design that would detect it.

CUSTOMERS	VISIT LIFT	RELATIVE	Z-STAT	REVENUE / CUST.
64,000	+6.1pp	+57%	22	+$0.60

Reading the Effect

Emailed customers visit the site at **16.7%** against **10.6%** for the held-back control, a **+6.1 percentage-point** causal lift (a 57% relative increase). The effect carries downstream: conversion rises from 0.57% to 1.07%, and average revenue per customer lifts by **$0.60** over two weeks ($p < 0.001$). The split between the two creatives is not equal, which is the first sign that the right question is not whether to email but whom to email.

FIGURE 1 · VISIT RATE BY RANDOMIZED ARM



Site-visit rate for each arm against the held-back control (red line). Both emails lift visits; the mens creative lifts them by +7.7pp against +4.5pp for the womens creative, a heterogeneous effect the average hides.

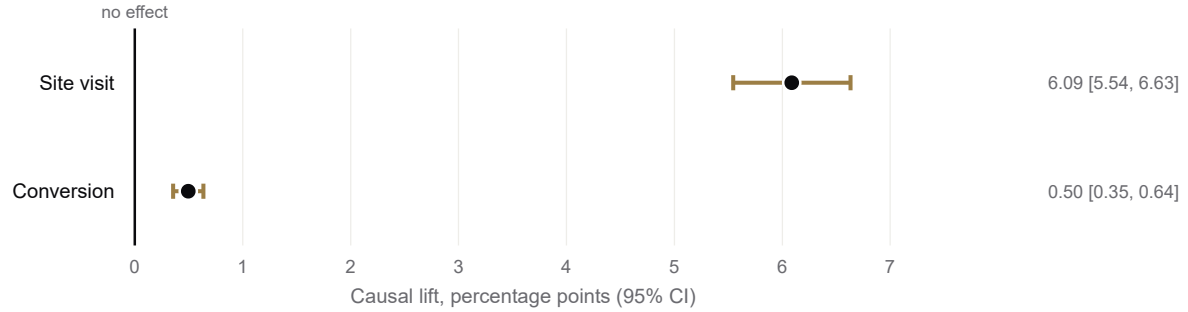
> **THE POINT**
> Randomization is what turns a correlation into a cause. Because assignment was random, the gap between the arms is the campaign's effect and nothing else, which is why the experiment, not the model, is the strongest evidence in analytics.

Sources. 1. Design and analysis of randomized controlled experiments: R. Kohavi, D. Tang & Y. Xu, **Trustworthy Online Controlled Experiments** (Cambridge, 2020). https://experimentguide.com/ 2. Data: K. Hillstrom, MineThatData E-Mail Analytics And Data Mining Challenge, 64,000 customers randomized to mens-email / womens-email / no-email, 2008. https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html

Is the Lift Real, or Noise?

With 64,000 customers the effects are estimated tightly. The visit lift is **+6.09pp** with a 95% confidence interval of **[5.54, 6.63]**, far from zero (z = 22); conversion lifts **+0.50pp** [0.35, 0.64]. A confidence interval that clears zero by this margin is what separates a real effect from sampling noise, and it is why sample size is a design decision, not an afterthought.[3]

FIGURE 2 · TREATMENT EFFECTS WITH 95% CONFIDENCE INTERVALS

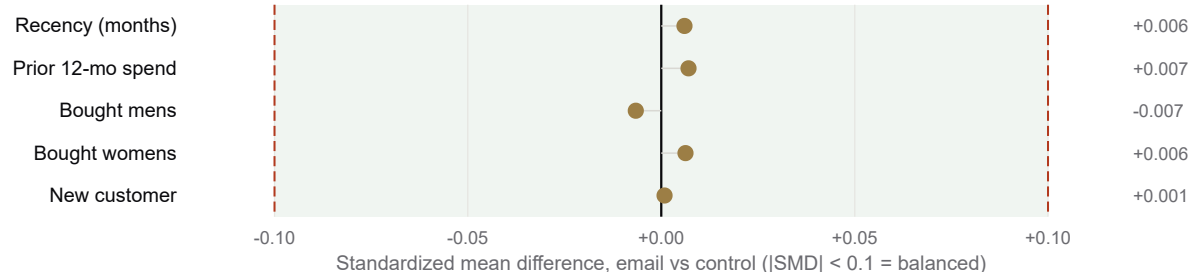


Causal lift in percentage points for each outcome, with 95% intervals. Both sit entirely right of the no-effect line, significant at the 5% level.

Is It Really Causal? The Balance Check

The causal claim rests on the arms being comparable before treatment. A balance check confirms it: across every pre-treatment covariate the standardized mean difference between emailed and control customers stays under **0.01**, well inside the 0.1 rule of thumb for balance.[4] Adding those same covariates to the model barely moves the estimate (6.09pp to 6.10pp), which is exactly what should happen when randomization has already removed confounding: adjustment becomes a check, not a correction.

FIGURE 3 · COVARIATE BALANCE, EMAIL VS CONTROL



Standardized mean difference for each pre-treatment covariate. Every point sits inside the plus/minus 0.1 balance band, confirming the randomization held and the raw difference is causal.

Who Responds, and Sizing the Next Test

The mens creative lifts visits by **+7.7pp** against **+4.5pp** for the womens creative, so the return sits in targeting the treatment rather than blanket sending, the case for an uplift model over a response model. The design is checkable too: at the control visit rate of 10.6% and 21,306 per arm, the test detects a lift as small as **0.8pp** (8% relative) at 80% power and 5% significance, which is how you size the sample before spending on the next campaign.[5]

Limitations & Method

Honest scope. A two-week window measures the short-run effect, not repeat purchase or unsubscribe cost, and one experiment does not tune a targeting policy alone. The transferable discipline is the chain: randomize, estimate with an interval, prove balance, read heterogeneity, size the next test. **Method:** two-proportion z-tests and a Welch t-test; HC1 robust regression for covariate adjustment; standardized mean differences for balance; a normal-approximation power calculation for the minimum detectable effect.

Sources. 3. Estimation and interpretation of causal effects under randomization: G. Imbens & D. Rubin, **Causal Inference for Statistics, Social, and Biomedical Sciences** (Cambridge, 2015). https://doi.org/10.1017/CBO9781139025751 4. Standardized mean difference as a balance diagnostic (|SMD| < 0.1): P. Austin, "Balance diagnostics," **Stat. in Medicine** 28(25), 2009. https://doi.org/10.1002/sim.3697 5. Power and minimum-detectable-effect for two proportions: J. Cohen, **Statistical Power Analysis for the Behavioral Sciences** (2nd ed., 1988). https://doi.org/10.4324/9780203771587