Experiments, A/B Testing, and Causal Evidence

From Prediction to Cause

Dr. Jose Mendoza, Academic Director and Clinical Associate Professor
Version 1.0 · July 2026

Chapter Information

ABSTRACT

This chapter builds the causal branch of marketing evidence, with StyleCraft's Drop Alert rollout decision as its running problem. It establishes what causal claims require: the counterfactual, potential outcomes at intuition level, and the fundamental problem that only one of a customer's two possible outcomes is ever observed. Selection bias is then named as the reason observational comparisons flatter targeted campaigns, and decomposed arithmetically on a record where the treated group's composition accounts for two-thirds of the apparent effect. Random assignment follows, with control design, the unit of randomization, preregistration as the experimental form of the analytic specification, statistical power, intention-to-treat analysis, and the checks that grade the assignment before the analysis grades the campaign. The chapter then measures lift with its interval, prices it against a break-even line computed from the program's own dose and economics, catalogs five testing failures, bounds quasi-experimental inference, and closes on the evidence hierarchy.

KEYWORDS

causal inference; counterfactual; selection bias; random assignment; A/B testing; preregistration; statistical power; lift; experimentation; difference-in-differences

VERSION AND DATE

Version 1.0 · July 2026 · Language: English (United States)

SUGGESTED CITATION

Mendoza, J. (2026). Experiments, A/B testing, and causal evidence. In *Applied business analytics for marketing decision-making: Business analytics and data visualization* (Chapter 11, Version 1.0) [Open educational resource]. CC BY 4.0.

of GitHub, Inc. Product names are used for identification only and do not imply endorsement. StyleCraft Collective is a fictional company created for instruction.

COMPANION REPOSITORY
Datasets, notebooks, and figure sources for this chapter: Applied Business Analytics companion repository on GitHub

Chapter Learning Objectives

By the end of this chapter, students should be able to:

- Explain why marketing decisions are causal questions, state the fundamental problem of causal inference in the potential-outcomes frame at intuition level, and identify the missing counterfactual behind any claim that a campaign "worked."
- Define selection bias, explain why observational comparisons systematically flatter well-targeted marketing programs, state what determines the direction of the bias, and decompose an observed group difference into a treatment component and a composition component on a worked example.
- Explain what random assignment buys — comparability in expectation on all characteristics, measured and unmeasured — state the additional conditions an unbiased effect estimate requires, and design a treatment and control comparison for a stated marketing decision.
- Choose a unit of randomization appropriate to a marketing intervention, state the contamination risk that follows from choosing a unit smaller than the unit the treatment actually reaches, and explain why effect estimates and standard errors must respect the assignment grain.
- Write a preregistration for a marketing experiment — decision, primary metric, arms, treatment dose, unit of randomization, sample size, duration, decision rule, and stopping rule — as the experimental form of the analytic specification of Section 2.5, and confirm that every declared guardrail is computable from the certified data.
- Explain statistical power and the relationship between detectable effect size and required sample size, and compute the sample size implied by a stated minimum detectable lift, baseline rate, significance level, and power target.
- State the intention-to-treat principle, build an assignment-flow table from randomization through outcome observation, and explain why post-assignment filtering recreates the selection the design was built to remove.
- Interpret a covariate balance check and a sham-split analysis-path check correctly, distinguish both from a genuine end-to-end A/A test, and state what each failure mode of the assignment machinery would look like in their output.
- Compute absolute and relative lift for a conversion experiment, construct a confidence interval for a difference in proportions and a separate interval for a ratio of proportions, and report the result in percentage points, in relative terms, and in dollars.

- Distinguish statistical from practical significance, define a program's treatment dose, compute a break-even lift from that dose and the program's contribution, and state a rollout decision rule that compares the full interval against that line under stated assumptions.
- Recognize and prevent the five standing testing mistakes — peeking, multiple comparisons, mid-test changes, novelty effects, and contamination — and identify each from the symptoms it leaves in a test report.
- Describe pre/post with a comparison group and difference-in-differences at concept level, state the parallel-trends assumption these designs require, and explain what they can and cannot establish when randomization is unavailable.
- Place a piece of marketing evidence on the evidence hierarchy, state the strongest sentence each rung licenses, and state the execution-quality condition the hierarchy assumes.
- Audit an AI-drafted experiment design and analysis with the five-point audit of Section 8.11 plus the experimentation supplement — registration precedes analysis, assignment integrity verified, inference honesty checked — predicting the estimate before running it, and documenting the audit per Appendix D.
- Assemble a rollout recommendation a chief marketing officer can act on and a finance partner can challenge: registration, assignment certificate, lift with interval in decision units, break-even comparison with its sensitivities, extrapolation assumptions named, and the guardrails under which the rollout would be reversed.

Chapter 10 ended with a forecast that could be graded. The buy plan's number arrived with a method, an opponent it had beaten, a sealed-holdout grade in dollars and percentages, a prediction interval that stated its own ignorance, and an override log that made the analyst's private knowledge of the calendar into a public, reviewable input. That deliverable answers one class of question completely: what will happen if StyleCraft keeps doing what it is doing. It answers no version whatever of the question the same meeting asks thirty seconds later — what would happen if we did something different — and the reason is structural rather than technical. Every method in Part II learned from a single history in which StyleCraft did exactly one thing. The forecast can report that suburban revenue stepped up in the weeks after the Greenwich opening; it cannot report what those weeks would have contained had the store never opened, because that version of the fall does not exist in any file and never will. The gap between the two statements is not a gap in the data's size or the model's sophistication. It is a gap in the kind of evidence that observation can produce at all, and it is the reason this chapter exists. Marketing decisions are actions — send it, don't send it, spend here, spend there — and an action is a claim about a world that has not happened yet. This chapter builds the only instrument in the guide that manufactures evidence about such worlds on purpose: the controlled experiment, in which the analyst does not wait for a comparison to appear in the data but creates one, by deciding at random who gets the treatment and who does not. The chapter's arithmetic is the simplest in Part II — two proportions and their difference — and its discipline is the strictest, because an

experiment's credibility is spent almost entirely in the choices made before any number is computed, and can be destroyed afterward by a single unplanned look at the results.

CONCEPT

What This Chapter Is Really About

Every chapter of Part II has produced a number that describes the world StyleCraft already lives in. The segments describe who its customers are; the regression describes which characteristics accompany higher spend; the churn model describes who is likely to leave; the forecast describes where the series is heading. Not one of those numbers, however carefully earned, answers the question that actually precedes a marketing budget: if we act, what changes? The obstacle is not that the models are weak. It is that the data records a world in which StyleCraft's choices were made by StyleCraft — targeted at the customers judged most likely to respond, timed to the seasons with the most demand, aimed at the stores already performing — so that every apparent effect in the record arrives fused to the reasons it was applied where it was. Untangling the two by thinking harder is impossible in principle, because the comparison the question needs is a version of the same customers, in the same season, who did not receive the campaign, and that version was never recorded.

The experiment's contribution is to stop trying to untangle and start manufacturing: if a chance mechanism decides who is treated, then the treated and untreated groups differ, on average, in exactly one respect — the treatment — and their difference in outcomes estimates the effect with nothing else fused to it. That single structural move is the whole chapter, and everything else in it exists to protect the move from being undone. The preregistration protects it from the analyst's own hindsight. The integrity checks protect it from a broken randomizer. The intention-to-treat rule protects it from a delivery failure quietly rebuilding the selection. The power calculation protects it from a test too small to see the effect it was built to find. The stopping rule protects it from the peeking that manufactures winners out of noise. The break-even line protects the result from being read as a business case it does not support. And the evidence hierarchy protects the rest of the guide's output from being described in words only an experiment has earned. An experiment is the most trustworthy instrument in this book, and the most fragile.

Source: Course concept developed for this guide, informed by Kohavi et al. (2020) and Imbens and Rubin (2015).

11.1 Marketing Decision Context: Did the Campaign Work, or Did the Customers?

The Comeback Edit ranked customers by risk; the buy plan committed inventory against a season. The decision that opens this chapter spends no money on customers at all until someone settles an argument about a number that has already been computed, twice, by two people who do not trust each other.

The number belongs to the Drop Alert. StyleCraft's merchandising rhythm is the weekly drop — a Thursday release of new styles, announced across email, social, and the app — and eight

months ago the CRM team launched a variation on it: an SMS and app-push notification giving recipients a twelve-hour early-access window to the week's drop before the general announcement. One alert, one drop. The program costs roughly twelve cents per enrolled customer per alert once messaging fees and the operations of staging an early window are counted, and it has been running on the customers the CRM team judged most likely to use it — the opted-in, the app-installed, the loyalty-enrolled. The numbers have looked extraordinary. Customers who received the Drop Alert converted at 10.1 percent in the two weeks following a drop; customers who did not converted at 4.8 percent. The difference is 5.3 percentage points, which the CRM team's deck renders, correctly as arithmetic, as a lift of more than 110 percent.

On the strength of that deck, the chief marketing officer has proposed a reallocation for the fourth quarter: four hundred thousand dollars moved out of paid social and into an expansion of the Drop Alert to the full SMS-eligible base, beginning with the November drop calendar. The logic is not foolish. A program that doubles conversion at twelve cents a message is, if the number means what it appears to mean, the best-performing marketing StyleCraft owns, and the holiday quarter is precisely when it should be scaled. The CRM calendar for November and December locks in three weeks, because creative, merchandising, and messaging operations all schedule against it, which converts an interesting analytical question into a dated one.

The finance partner has blocked the reallocation, and her objection is not about the arithmetic. It is that the two groups were never comparable. The customers who received the Drop Alert were selected — deliberately, and for good operational reasons — because they were the most engaged customers StyleCraft has: app users, SMS opt-ins, the loyalty tiers Chapter 6's segmentation named the Urban Loyal Core. The customers who did not receive it are disproportionately the Suburban Occasion and New/At-Risk populations, who would convert at a fraction of the Core's rate if StyleCraft sent them nothing at all. Her question is the one this chapter answers: how much of the 5.3 points is the Drop Alert, and how much is the difference between the two kinds of customer who happened to land on either side of the comparison? She will fund the expansion, at scale, if someone can show her a number that is not contaminated by who was chosen.

So the VP of Marketing has borrowed the analyst — still you — and the commission is different in kind from every commission in Part II. Nothing in the existing data will settle the argument, because it records exactly one world: the one in which StyleCraft sent the Drop Alert to its best customers. What settles it already exists. Six weeks ago, anticipating this fight, the CRM team ran a proper test: ten thousand eligible customers were listed, a seeded draw selected exactly five thousand to receive one Drop Alert for the drop in their assignment week, the remaining five thousand received the standard announcement for the same drop, and conversion within the following fourteen days was declared in advance as the deciding metric. That file is the ab_test table, and what this chapter builds from it is the campaign's causal effect on conversion — a lift with an interval, priced at a stated dose against the break-even line that makes the rollout worth funding, with an honest account of what the experiment cannot say about a rollout differing from the test in population, duration, and dose.

Two features of the commission shape everything after. The first is that the answer is knowable in advance, and this is deliberate: the design note published with the course data dictionary states the effect the generator planted, a lift of 1.8 percentage points — large enough to be detectable at the planned sample size, small enough that a sloppy analysis will miss it or exaggerate it. That is the chapter's verification theme, and the analysis therefore grades itself. The second is that the observational number is not a lie. Both rates are real and both were computed correctly. The 5.3-point difference is a true fact about StyleCraft's records and a false answer to the CMO's question, and holding both in mind at once is the discipline this chapter installs.

The receipts lock in three weeks. Sections 11.2 and 11.3 explain why the true fact is the wrong evidence. Sections 11.4 and 11.5 build the design that produces the right evidence, size it, and audit it. Sections 11.6 and 11.7 measure the effect and price it against the program's dose. Section 11.8 catalogs the ways a sound measurement is destroyed after the fact. Sections 11.9 and 11.10 handle the decisions no chance mechanism can reach and place all of Part II's output on the ladder of claims. Section 11.11 turns to the assistant. The labs in Section 11.12 recover the planted lift and manufacture a false winner on purpose, and Sections 11.13 through 11.15 rehearse the meeting where an interval becomes a rollout.

11.1.1 Opening Case Questions

Keep these questions in mind while reading, and return to them after completing the labs.

- The CRM team's 5.3-point difference and the experiment's answer will disagree by a factor of roughly three. Before reading further, write one sentence naming the mechanism that produces the gap, and one sentence explaining why adding control variables to a regression on the observational data would not reliably close it.
- The finance partner asked for "a number that is not contaminated by who was chosen." State, in your own words, what property of the experiment's assignment procedure delivers exactly that, and what the property guarantees about characteristics of customers that StyleCraft has never measured.
- The experiment measured conversion over fourteen days among SMS-eligible customers during a non-holiday period, at a dose of one alert. The rollout would run for a quarter, across the holiday calendar, at a higher alert frequency. List three reasons the experimental estimate might not transfer, and note which of them this chapter names.
- The Drop Alert costs twelve cents per customer per alert, and the test delivered one alert per treated customer. Before Section 11.7 tells you, work out what additional piece of information you would need to decide whether a 1.8-point lift is worth having, and write the arithmetic you would perform once you had it.
- Suppose the experiment had been run at two alerts per treated customer instead of one, with everything else unchanged. State what that would do to the cost per assigned customer, to the break-even lift, and — this is the part worth thinking about — to the recommendation the same measured interval would support.

11.2 The Fundamental Problem of Causal Inference

The opening case turns on a question that sounds empirical and is, in a specific and permanent sense, unanswerable as posed: did the Drop Alert cause the conversion? This section explains why the question cannot be answered about any individual customer, ever, by anyone, with any amount of data — and why that impossibility, properly understood, is the design brief for everything the rest of the chapter builds.

Start with one customer. C004471 received the Drop Alert on a Thursday in September and bought a dress the following Tuesday. Did the Drop Alert cause the purchase? The honest form of the question is comparative: what would C004471 have done if she had not received it? There are exactly two possibilities to compare — the outcome under treatment and the outcome under no treatment — and this pair is what the modern causal literature calls her potential outcomes (Rubin, 1974; Holland, 1986). Her causal effect is the difference between the two. And the difficulty, which no technology will remove, is that the world assigned her one of them. She received the alert; she bought; the version of C004471 who did not receive the alert is not in the file, not recoverable from the file, and does not exist. Her second potential outcome is not missing in the sense that Chapter 4's defect catalog uses the word, where a value was recorded badly or lost in transit. It is missing in the sense that it never happened.

> **DEFINITION**
>
> **Counterfactual and Potential Outcomes**
>
> For a unit that could be exposed to a treatment or not, the potential outcomes are the pair of results that would occur under each exposure: the outcome under treatment and the outcome under no treatment. The unit's causal effect is the difference between the two. The counterfactual is whichever of the pair did not occur — the outcome in the world that was not realized. Because a unit is exposed either to the treatment or to its absence but never to both, exactly one potential outcome is observed and the other is permanently unavailable, which is why causal effects are estimated by comparison across units rather than measured within them.
>
> Source: Adapted from Rubin (1974) and Holland (1986).

In other words, every causal claim in marketing is a claim about something that did not happen. "The campaign lifted conversion by 1.8 points" is shorthand for "conversion among these customers was 1.8 points higher than it would have been in a version of the same fortnight in which they received nothing," and the second half of that sentence is a description of a world nobody observed. Holland (1986) gave the situation its standing name — the fundamental problem of causal inference — and the name is worth taking literally. It is fundamental in that it is a feature of the logic of causation rather than a limitation of current measurement; it is a problem in that it makes the quantity of interest structurally unobservable; and it belongs to inference, not to data collection, because the response is not to collect the missing outcome but to construct a defensible estimate of it.

DEFINITION

Causal Inference

Causal inference is the estimation of what an outcome would have been under an alternative exposure — the counterfactual — using observed data plus assumptions or design features that make some observed group a credible stand-in for the unobserved one. Its subject is the effect of an intervention rather than the pattern of an association: it answers what would change if we acted, not what accompanies what in the record. Its credibility rests on the plausibility of the stand-in, which is established by research design rather than by the size of the dataset or the sophistication of the model fitted to it.

Source: Adapted from Holland (1986) and Imbens and Rubin (2015).

The closing sentence of that definition is the chapter's hinge, and it deserves to be stated at full force because it contradicts an intuition that four chapters of modeling have carefully built. Nothing in Chapters 7 through 10 improves as evidence about causes when the dataset grows. A regression on eighty thousand order lines estimates its coefficients more precisely than a regression on eight hundred, and precision is not the currency causation trades in: an estimate can be pinned to three decimal places and still be an estimate of the wrong quantity. The missing counterfactual is not estimated more accurately by more rows of the same kind, because more rows of the same kind describe the same single world in more detail. What changes the answer is a change in how the comparison group came to be a comparison group — which is a decision about design, made before data exists, and the reason this chapter's most consequential material sits in Sections 11.4 and 11.5 rather than in the arithmetic of Section 11.6.

The practical move that rescues the situation is a shift of ambition, from the individual to the group. C004471's personal effect is unknowable, and StyleCraft does not need it. What the CMO's decision requires is the average effect across the customers the program would reach: the difference between the conversion rate this population would show if all of them received the Drop Alert and the rate the same population would show if none of them did. Both of those quantities are still counterfactual — no population is simultaneously fully treated and fully untreated — but the average has a property the individual effect does not. If StyleCraft can find or construct two groups of customers that are alike, on average, in every respect that matters for conversion, then one group's realized outcome can stand in for the other group's missing one, and the difference between their observed rates estimates the average effect. The whole apparatus of experimentation is a response to the phrase "alike, on average, in every respect that matters." Section 11.3 shows what happens when the groups are not alike, which is the ordinary condition of marketing data. Section 11.4 introduces the one procedure known to deliver the condition without requiring anyone to list what matters in advance.

11.3 Selection Bias: Naming the Confound Problem

Section 7.5 named the confounder and demonstrated it on StyleCraft's own calendar, showing that email volume and revenue correlate because the merchandising season drives both. It ended by identifying three responses — compare within the confounder, control for it in a model, or change the design so the confounder cannot operate — and it explicitly deferred the third to this chapter. This section takes up the deferral by giving the problem its causal name, showing exactly how much damage it does to the opening case's number, and explaining why the first two responses, which Chapter 7 taught, are insufficient here in a way that is not a matter of effort.

DEFINITION

Selection Bias

Selection bias is the difference between an observed group comparison and the causal effect it is taken to estimate, arising because the units that received the treatment differ systematically from those that did not in ways that themselves affect the outcome. The comparison then mixes two quantities that cannot be separated within it: the effect of the treatment and the effect of the differing composition of the groups. Selection bias is a property of how the groups came to exist, not of how carefully the outcome was measured, and it is unaffected by the size of the dataset. Its direction follows the selection mechanism rather than being fixed in advance.

Source: Adapted from Angrist and Pischke (2009) and Imbens and Rubin (2015).

In other words, the confounding of Section 7.5, seen from the causal side, becomes a question about who ended up in which group and why. And marketing manufactures selection bias as a matter of professional competence, which is what makes it so treacherous here. A CRM team that sends its best campaign to its most responsive customers is not making an error; it is doing its job, and doing it well. But the same competence guarantees that the treated group in the resulting record is composed of people who would have converted at a higher rate than the untreated group even if the campaign had never been built.

The direction of the resulting distortion deserves more care than marketing folklore usually gives it. When high-propensity customers are preferentially treated, selection commonly inflates the observed program difference, and that is the case StyleCraft's Drop Alert presents. Other targeting rules produce the opposite. A win-back program aimed at lapsed customers, a save offer aimed at accounts already showing warning signs, a discount aimed at the weakest markets, an onboarding sequence aimed at customers with no purchase history — every one of these treats the customers least likely to convert, and the observational comparison then understates or even reverses the program's true effect. The direction of selection bias follows the selection mechanism, and an analyst who assumes it always points upward will misread half the programs she evaluates. What is general is not the sign but the fusion: the record cannot separate treatment from selection on its own. Gordon et al. (2019) put that point on the record at industrial scale, running large advertising experiments and then attempting to reproduce their answers using the observational methods practitioners rely on when experiments are unavailable, with access to a

volume of behavioral covariates far beyond what most marketing organizations possess; the observational estimates diverged from the experimental ones substantially and unpredictably, in both directions and by large margins.

The Drop Alert's own numbers make the mechanism arithmetically visible, and the arithmetic is small enough to do by hand — which is where Lab 11.1 begins. Take the eight months of observational history and split it not into two groups but into four, by crossing the treatment with the one confounder StyleCraft happens to have measured: whether the customer was engaged, meaning app-installed and loyalty-enrolled, at the time of the send. The CRM team's targeting shows up immediately in the composition. Of the sixteen thousand customers in the comparison, eight thousand received the Drop Alert and 75 percent of them were engaged; eight thousand did not receive it and only 25 percent of them were engaged. And the outcomes, arranged this way, tell a different story than the headline did.

Table 11.1

The Drop Alert's observational record, arranged two ways

Group	Customers	Conversions	Conversion rate
Received, engaged	6,000	708	11.80%
Received, not engaged	2,000	96	4.80%
Received, all	**8,000**	**804**	**10.05%**
Not received, engaged	2,000	200	10.00%
Not received, not engaged	6,000	180	3.00%
Not received, all	**8,000**	**380**	**4.75%**

Read the bold rows and the deck's number appears: 10.05 percent against 4.75 percent, a difference of 5.30 percentage points, a relative lift of 112 percent. Read the unbolded rows and something more interesting is visible, twice. Among engaged customers, the treated converted at 11.80 percent and the untreated at 10.00 percent — a within-stratum observational difference of 1.80 points. Among customers who were not engaged, 4.80 against 3.00 — a within-stratum observational difference of 1.80 points. The two strata agree, which is the designed signature of a record in which the treatment did the same thing to both kinds of customer.

What the agreement licenses is narrower than it looks, and the distinction matters enough to state before the decomposition rather than after. Under this miniature's designed assumption that engagement is the only confounder, that it has been measured without error, that no unmeasured selection remains, and that treated and untreated customers are otherwise exchangeable within each stratum, standardizing on the treated group's engagement mix recovers an adjusted

difference of 1.8 points. That adjusted difference is what the observational record can produce. It is not yet a causal effect, and the reason is not arithmetic but architectural: the four assumptions above are assumptions, they were installed by the example's designer, and nothing in the record can check them. The randomized experiment of Section 11.4 is what establishes the causal interpretation without requiring any of them.

With that qualification in place, the decomposition does its work. The remaining 3.5 points is composition, and it decomposes exactly. The treated group was 75 percent engaged and the untreated group 25 percent engaged, a difference in mix of 50 percentage points; engaged customers converted 7 points higher than unengaged ones absent any treatment (10.00 against 3.00); and half of a seven-point gap is 3.5 points of pure difference-in-kind, contributed by nothing but who was chosen. Two-thirds of the headline was the CRM team's targeting skill being read back to it as campaign performance. An equivalent way to see it, and the one worth carrying into the lab: ask what the treated group would have converted at if the campaign had never existed. Its own mix, at untreated rates, gives $0.75 \times 10.00 + 0.25 \times 3.00 = 8.25$ percent — and the treated group's realized 10.05 percent minus that constructed 8.25 percent is 1.80 points, the adjusted difference, while the constructed 8.25 percent minus the untreated group's 4.75 percent is 3.50 points, the composition component. The two quantities are additive, they are both real, and only one of them is even a candidate answer to the CMO's question.

It is tempting to conclude that the problem is solved: stratify by engagement, average the within-stratum differences, and report 1.8 points. Chapter 7 taught exactly that move, and here it worked perfectly. It worked, however, for a reason that will almost never hold in practice — the example was built with a single confounder, that confounder was measured, and it was the only one operating. Real targeting decisions are made on dozens of measured variables and at least as many unmeasured ones: the CRM manager's judgment about which customers "get" the format, the historical accident of which cohorts were enrolled during which promotion, the customer's own decision to install the app, which is itself a symptom of purchase intent that no column records. Controlling for a confounder requires having measured it, and having measured it requires having thought of it. That is the asymmetry that makes observational causal claims in marketing so unreliable: the analyst can enumerate what she controlled for, and can never enumerate what she failed to. No observational model can guarantee separation of treatment and selection without assumptions about the variables and mechanisms responsible for assignment. The lesson is not that statistical adjustment is useless. It is that adjustment can correct only the imbalances the analyst can name, while the design of Section 11.4 corrects all of them at once, including the ones nobody has thought of yet.

CONCEPT

The Comparison Is the Deliverable

By the end of this chapter the arithmetic will occupy four lines of code and the argument will occupy four pages, and the imbalance is the correct one. Two proportions and their difference is not hard mathematics; every spreadsheet in the building can compute the 5.3 points that started the argument. The question a marketing analyst is actually paid to answer is not what the difference between two groups is but whether that difference means what the meeting will take it to mean, and that question is settled entirely by how the two groups came to be groups.

Hence the working rule this chapter enforces from here forward, in the labs, in the deliverable, and in the exercises: before quoting any comparison as evidence about a decision, state in one sentence how each side of it was selected, and state what would have made the two sides differ in outcome even if the treatment had never existed. If the answer to the second question is "nothing, because a chance process assigned them," the comparison is evidence about a cause. If the answer is a list of plausible characteristics, the comparison is a description — sometimes a useful one — and the words used to report it must fall back accordingly, per Section 11.10's ladder.

Source: Course concept developed for this guide, informed by Angrist and Pischke (2015) and Gordon et al. (2019).

11.4 Random Assignment and Experiment Design

Section 11.3 ended on an asymmetry: adjustment can only correct the imbalances an analyst can name. This section introduces the procedure that corrects all of them without requiring any to be named, which is the single most important idea in the chapter and among the least intuitive, because its power comes from an act that looks, at first, like an abdication of judgment.

The procedure is to decide who receives the treatment by a chance mechanism — a coin, a random number generator, a seeded draw — with the decision made in advance and with no input whatsoever from anyone's knowledge of the customers. The mechanism admits two standard forms, and the difference between them is not pedantry, because it determines what the arm sizes are allowed to be. Under Bernoulli randomization each unit is assigned independently with a fixed probability, so the realized arm sizes are themselves random and need not come out equal. Under complete randomization a fixed number of units is drawn at random from the eligible list, so the arm sizes are exactly what the design declared. StyleCraft's experiment used the second, and the registration says so: from the ten thousand eligible customers, a seeded random draw selected exactly five thousand for treatment, and the remaining five thousand formed the control arm. That choice is what makes the arm-size check of Section 11.5 meaningful, because a realized split of 5,000 and 4,780 under complete randomization is not bad luck — it is evidence that something happened after assignment. Nobody chose the engaged customers for the treatment arm. Nobody chose anybody for anything.

DEFINITION

Random Assignment, Treatment, and Control

Random assignment is the allocation of units to experimental conditions by a chance mechanism that is independent of every characteristic of the units, so that each unit's condition is statistically unrelated to who that unit is. The treatment group receives the intervention under test; the control group receives the existing experience, no intervention, or a declared alternative, and supplies the estimate of what the treatment group's outcome would have been in the absence of the treatment. Because assignment is independent of unit characteristics, the two groups are comparable in expectation on every characteristic — recorded or unrecorded, known or unsuspected.

The difference in outcomes is an unbiased estimate of the average effect of the treatment under conditions the design must also satisfy: assignment was implemented as planned, outcomes are observed consistently across arms, units are analyzed according to their original assignment, no unit's outcome is affected by another unit's assignment, and the treatment itself is well defined at a stated dose. Randomization delivers comparability; those five conditions deliver the estimate.

Source: Adapted from Fisher (1935) and Imbens and Rubin (2015).

The phrase carrying all the weight in the first paragraph of that definition is "recorded or unrecorded." Randomization does not make the two groups identical, and any student who checks will find that StyleCraft's treatment arm contains slightly more Gen Z customers, slightly fewer Backstage members, and a marginally different mean of prior ninety-day spend than its control arm. What randomization guarantees is that these differences arise only from the draw, are as likely to fall one way as the other, shrink as the arms grow, and — crucially — are unrelated to anything about the customers, including the twenty characteristics StyleCraft never measured and the five it does not know exist. This is why the design corrects for confounders nobody has thought of. The chance mechanism does not need to know what it is balancing. It balances everything, in expectation, because it consults nothing.

Two consequences follow, and both are routinely misunderstood in practice. The first is that randomization is a statement about the procedure, not about the resulting sample. A single experiment can draw an assignment in which the arms differ noticeably on some characteristic; the guarantee is that this is improbable and undirected, not impossible. This is exactly why the checks of Section 11.5 exist: not to "prove" the randomization worked, which no check can do, but to catch the two situations that actually occur in practice — a randomizer that was never random (assignment correlated with signup order, customer identifier range, or the hashing of a field that encodes geography) and an assignment that was random but was subsequently corrupted by an operational failure such as a targeting rule quietly re-filtering the treatment list. The second consequence is that randomization buys comparability and nothing else. It does not make the outcome metric well defined, and Chapter 3's four-element metric discipline still applies: conversion counted over what window, in what numerator, with what filters. It does not make the effect large, stable, or worth having. It does not make the experimental population representative of the rollout population — a point Section 11.13 will spend a full argument on.

And it does not protect the result from the analytical mistakes of Section 11.8, all of which are committed after the assignment is complete. Random assignment removes exactly one problem, completely, and leaves every other problem in the room.

Designing the comparison, then, is mostly a matter of specifying what the control arm receives, and this is where inexperienced designs go wrong first. "Control" does not mean "nothing happens." It means the arm receives the alternative against which the treatment must prove itself, and that alternative should be the thing the business would otherwise do — which for StyleCraft is the standard drop announcement, not silence. A test of Drop Alert against silence would answer a question nobody asked (does messaging beat not messaging?) and would report a lift inflated by the entire value of ordinary communication. The control arm is the baseline, in the direct lineage of the discipline Section 2.7 introduced and Section 8.7 operationalized: the opponent the treatment must beat, chosen to represent the decision's real alternative rather than a strawman. When the real alternative is genuinely nothing — an incremental advertising test where the decision is whether to spend at all — then silence is the right control, and the arm is usually called a holdout, with the ethical questions Section 11.15 takes up.

The second design decision is the one most often made by default and most expensive to get wrong: what, exactly, is being randomized.

DEFINITION

Unit of Randomization

The unit of randomization is the entity to which the chance mechanism assigns a condition — a customer, a session, a device, a store, a geographic market, or a time period. It determines what the experiment's arms are made of and what the sample size actually is for purposes of power. Outcomes may be recorded at a finer grain than the assignment, but treatment variation exists only at the assignment grain, so effect estimates and their standard errors must respect the randomization unit or its cluster structure. The unit should be at least as large as the unit through which the treatment can reach a person, because any pathway by which one unit's treatment influences another unit's outcome undermines the comparison.

Source: Adapted from Kohavi et al. (2020).

The second sentence of that definition closes a hole that swallows a surprising number of commercial test reports. An experiment that randomized ten thousand customers does not acquire more evidence because those customers generated forty thousand orders. The orders are more observations of the same ten thousand randomizations, and an analysis that counts them as independent shrinks the interval to a width the design never earned — the grain slip Section 11.11 names among the assistant's standing failures. More order-level rows are welcome as description. They are not more randomized evidence.

The rule of thumb hidden in the definition's last sentence is that the unit must be large enough to contain the treatment's spillover. Table 11.2 lays out the marketing cases and what

each choice buys and risks; it is a design checklist, and the labs and exercises return to it whenever an experiment must be specified rather than analyzed.

Table 11.2

Choosing the unit of randomization

Unit	Appropriate when	What it buys	What it risks
Session or visit	The treatment is a within-visit experience with no memory, such as a page layout tested on anonymous traffic	Very large sample sizes; fast results	The same person sees both conditions across visits, blurring the contrast and making a per-person outcome uninterpretable
Customer or account	The treatment is delivered to a person and the outcome accrues to that person, as with email, SMS, app push, or offers	A clean per-person outcome; the natural grain for CRM decisions	Spillover between customers who talk to each other; households sharing an account
Store	The treatment is a physical or operational change, such as assortment, staffing, or in-store signage	Realism; the treatment is applied as it would really be applied	Very small sample sizes — eighteen stores is eighteen units, not eighteen thousand customers — and correspondingly weak power
Geographic market	The treatment is broadcast and cannot be aimed at individuals, as with out-of-home, local TV, or regional media	The only feasible unit for untargetable media; captures total market effect	Few units, high variance between markets, long durations, and the need for designs like those in Section 11.9
Time period	Only one unit exists and the treatment can be switched on and off, as with a single store or a site-wide policy	Feasibility where nothing else is available	Trend and seasonality confound the on and off periods; weakest of the designs listed

StyleCraft's case resolves cleanly. The Drop Alert is delivered to a person, the conversion outcome accrues to that person, and the decision under debate is a per-customer program, so the unit is the customer — which is why the ab_test file has one row per customer, and why the analysis of Section 11.6 counts customers rather than messages, sessions, or orders. Note the discipline this imposes on the metric: a customer who converts three times in the window is one

conversion in a rate whose denominator is customers, per the definitional care of Section 3.5, and an analyst who quietly switches to counting orders has changed the question without saying so.

Contamination — the failure mode the unit choice exists to prevent — is treated in Section 11.8 with the other ways tests go wrong, because in practice it arrives late, as a discovery rather than a design flaw. But the design-time version of the question belongs here, and it is worth asking of every experiment before it launches: is there any path by which a treated unit's treatment can reach an untreated unit? For the Drop Alert the honest answer is that a small path exists — StyleCraft's most engaged customers follow the brand's social accounts and each other, and an early-access window that produces a visible sell-out of a style affects what remains available to the control arm hours later. The path is real, it is small at a five-thousand-customer scale against a national inventory, and the professional move is to record it in the registration as a known limitation with its expected direction (it would, if anything, depress the control arm's conversion and thereby inflate the measured lift) rather than to pretend it does not exist.

11.5 Preregistration, Power, and the Integrity of the Assignment

Random assignment makes an unbiased comparison available. It does not make an unbiased comparison inevitable, because the assignment is only half the experiment; the other half is a long series of analytical choices — which metric, which window, which customers, when to stop, which subgroups to look at — every one of which can be made after the outcomes are visible. An analyst who makes those choices while looking at the results can produce a favorable answer from a null effect essentially at will, without ever fabricating a number and usually without noticing what she is doing (Simmons et al., 2011). This section installs the discipline that closes the door, sizes the experiment that will test it, and then grades the machinery before the machinery grades the campaign.

DEFINITION

Preregistration

Preregistration is the practice of recording an experiment's design and analysis plan — the hypothesis, the primary outcome metric, the arms and their assignment, the treatment dose, the unit of randomization, the eligible population, the sample size and its justification, the planned duration, the analysis to be run, and the decision rule that maps the result to an action — in a dated, unalterable record created before the data are observed. Its function is to distinguish predictions from explanations: results specified in advance test a hypothesis, while results selected after inspection describe the data that generated them, and the two carry very different evidential weight even when the arithmetic is identical.

Source: Adapted from Nosek et al. (2018).

In other words, a preregistration is the analytic specification of Section 2.5 written for an experiment, with two elements added that Chapter 2 did not need: a stopping rule and a treatment dose. It carries the same virtues in the same order. It forces the decision to be named before the

analysis begins, so that the analysis is built for a decision instead of a decision being retrofitted to an analysis. It makes the deliverable auditable, because the record can be compared afterward against what was actually done. And it protects the analyst most of all, since an experiment whose result disappoints its sponsor is precisely the experiment whose method will be litigated line by line, and the only defense that survives that meeting is a document with a date on it.

Table 11.3 is the record StyleCraft's CRM team filed before the Drop Alert experiment launched, with a third column stating why each element must be settled in advance rather than chosen afterward. Two features of it are worth reading for specifically, because both are corrections to the way the draft of this chapter first stated them. The dose row exists because a program's cost per customer is meaningless until the number of exposures is fixed, and Section 11.7's break-even line is computed from that row rather than from a sentence in a slide. And every guardrail and secondary metric named in the table is computable from the certified ab_test file as the data dictionary defines it — the converted flag, the order_revenue column, and the pre-assignment characteristics — because a preregistered guardrail that cannot be calculated is not a minor gap in enrichment but a design defect that will be discovered at exactly the wrong moment. Unsubscribe and margin fields would make richer guardrails, and StyleCraft's registration would have declared them had the messaging and merchandising systems delivered them into the certified table on the experiment's schedule. They did not, so the registration declares what the certified table can measure and names the missing fields as an instrumentation obligation for the next test.

Table 11.3

The preregistration record for the Drop Alert experiment

Element	Declared value	Why it must be declared in advance
Decision	Whether to expand the Drop Alert to the full SMS-eligible base for the Q4 drop calendar	Names the action the evidence must support, per Section 2.3
Primary metric	Conversion: share of assigned customers with at least one order in the 14 days following assignment	One metric decides; a second promoted later is a multiple comparison (Section 11.8)
Arms	Treatment: Drop Alert early access. Control: standard drop announcement	Control is the real alternative, not silence
Treatment dose	One Drop Alert, for the single drop in the customer's assignment week	Fixes the cost per assigned customer, and therefore the break-even lift (Section 11.7)

Element	Declared value	Why it must be declared in advance
Assignment mechanism	Complete randomization: a seeded draw of exactly 5,000 of the 10,000 eligible customers into treatment	Fixes what the realized arm sizes should be, so a shortfall is a finding rather than noise
Unit of randomization	Customer	Fixes the analysis grain and the true sample size (Section 11.4)
Eligible population	SMS-opted-in customers with an account at least 30 days old	Defines whom the estimate is about, and therefore the rollout it licenses
Sample size	5,000 per arm	Justified by the power calculation below; fixed in advance so the stopping rule is meaningful
Duration	Assignment window of 14 days; outcome window of 14 days after each customer's assignment	Prevents mid-test extension in search of significance
Analysis	Intention-to-treat two-proportion comparison at the customer grain; absolute lift with a 95% interval on the difference; relative lift with a separate 95% interval on the ratio	Prevents the method from being chosen to fit the result
Decision rule	Expand if the lower bound of the interval on the difference exceeds the break-even lift of 0.43 points (Section 11.7)	Converts a statistical output into an action before anyone knows the output
Stopping rule	Analyze once, after the full outcome window closes for all assigned customers	The single most effective defense against peeking (Section 11.8)
Secondary metrics	Revenue per assigned customer; order count per assigned customer	Declared as secondary and exploratory, so they cannot become the headline
Guardrails	Revenue per assigned customer in the treatment arm must not fall below the control arm's by more than 5%	Names in advance the harm that would stop a rollout even if conversion won
Instrumentation obligation	Unsubscribe events and order margin to be delivered into the certified experimental table before the next test	A guardrail that cannot be computed is a design defect, not a missing nicety

Two entries in that table are earned by arithmetic rather than by declaration, and both belong to the same idea: an experiment that cannot see the effect it is looking for is not a cautious experiment but a wasted one.

DEFINITION

Statistical Power

Statistical power is the probability that an experiment will detect an effect of a specified size, at a specified significance level, given its sample size and the variability of its outcome — the chance that a real effect of that magnitude produces a statistically significant result rather than being missed. Power rises with sample size and with the size of the effect sought, and falls as the outcome grows more variable; because the required sample size grows roughly with the inverse square of the smallest effect worth detecting, halving the detectable effect roughly quadruples the sample needed. An underpowered experiment is not merely inconclusive: its significant results, when they occur, systematically overstate the effect's size.

Source: Adapted from Cohen (1988) and Kohavi et al. (2020).

The calculation that produced StyleCraft's five thousand per arm runs in one direction only, and it starts from the business rather than the statistics. The first input is the baseline conversion rate the control arm is expected to show, which StyleCraft reads off its own history at roughly 6 percent. The second is the minimum detectable effect: the smallest lift that would change the decision, which is a business judgment, not a statistical one, and which the CRM team set at 1.5 percentage points on the grounds that anything smaller would not justify the operational overhead of running early access across a national drop calendar. The third is the significance level, conventionally 5 percent, and the fourth is the power target, conventionally 80 percent. With those four numbers, the standard sample-size calculation for a difference in proportions returns 4,379 customers per arm; the team rounded to 5,000 for operational tidiness and to buy margin against the delivery failures that always occur. Code 11.7 makes the calculation executable rather than asserted, which is where students should verify every figure in this paragraph.

Read the calculation backward and it tells you something more useful than a number. At 5,000 per arm, StyleCraft has a 94.5 percent chance of detecting the 1.8-point effect the generator planted; at 1,000 per arm it would have had a 35.6 percent chance, meaning a two-in-three probability of running a perfectly designed experiment on a real effect and concluding there was nothing there. The inverse-square relationship stated in the definition is the fact worth memorizing, because it is what makes small effects expensive: detecting half a point instead of 1.5 requires not three times the sample but roughly eight and a half times it, near 36,800 customers per arm, which for many marketing questions exceeds the entire addressable base. Lewis and Rao (2015) turned this arithmetic into a general result about advertising, showing that because sales outcomes are so variable relative to advertising's true effects, the experiments needed to measure advertising returns with usable precision are far larger than the ones firms

typically run — which means many published advertising ROI figures are drawn from tests that never had the power to produce them.

With the size settled, three instruments grade the assignment rather than the campaign, and the order in which this chapter presents them is the order the deliverable follows. The first is the intention-to-treat rule, which decides who is in the comparison at all.

DEFINITION

Intention-to-Treat Analysis

Intention-to-treat analysis compares units according to the condition they were randomly assigned, regardless of whether the assigned treatment was actually delivered, received, or complied with. Its purpose is to preserve the property random assignment created: because assignment was independent of every unit characteristic, groups defined by assignment remain comparable, while groups defined by what happened after assignment — messages that reached a handset, customers who opened an app — are no longer randomized and reintroduce exactly the selection the design removed. The intention-to-treat estimate answers what the effect of offering the program is; the effect of the treatment actually received is a different quantity requiring stronger assumptions.

Source: Adapted from current trial-reporting guidance on retaining randomized participants in their assigned groups for the primary analysis (Hopewell et al., 2025) and Imbens and Rubin (2015).

This corrects a reflex the draft of this chapter encouraged and every experienced trialist would reject. When a 5,000/5,000 design reports 5,000 treatment customers and 4,780 control customers, the temptation is to conclude that 220 customers were "lost" and that the comparison is therefore invalid. The right reading is more precise and less alarming. A delivery failure is important operational evidence, and it is not by itself grounds for removing anyone from the primary comparison; removing the customers whose messages failed is precisely how an analyst converts a randomized comparison back into an observational one, because message delivery correlates with handset, carrier, tenure, and engagement. What the shortfall does demand is an explanation, and the instrument that produces it is a flow record.

Table 11.4

The assignment-flow record every experiment report must carry

Stage	Treatment	Control	What a gap at this stage means
Randomized	5,000	5,000	The design's declared allocation under complete randomization; any shortfall here is an assignment-code or list-construction defect

Stage	Treatment	Control	What a gap at this stage means
Message successfully delivered	reported from the messaging log	not applicable; control receives the standard announcement	Operational evidence about reach. It is reported, investigated, and never used to filter the primary comparison
Primary outcome observed	5,000	5,000	Outcome ascertainment. A gap here is the serious one: it means conversion could not be determined for some assigned customers
Included in the intention-to-treat analysis	5,000	5,000	Everyone randomized with an observed outcome, in the arm they were assigned to. This is the row the lift is computed from

Read the table as three different estimands wearing similar clothes. The intention-to-treat effect is the effect of assignment to the Drop Alert program, which is what a rollout decision actually needs, because a rollout also delivers imperfectly. The effect of treatment received is a different quantity, estimable only under assumptions about why delivery failed, and it is out of scope for this guide. And a missing outcome — a customer randomized whose conversion cannot be determined at all — is neither of those; it is a defect requiring investigation and, where it cannot be repaired, a sensitivity analysis that shows how the conclusion moves under the best and worst plausible assumptions about the missing customers. Only the last of these three licenses anyone to remove a randomized customer from a table, and even then it is reported as a deviation.

The second instrument compares pre-assignment characteristics across arms, and it is the one most often misused. The draft of this chapter implied that a prior-spend difference of more than a percent or two between arms would invalidate the simple randomized comparison. It would not, and reasoning that way inverts the logic of the design. Random assignment does not guarantee balance in every realized sample; chance imbalance can occur when the procedure is entirely valid, and baseline significance tests or arbitrary imbalance thresholds should not be used to decide retrospectively whether a valid randomization "worked" (Altman, 1985; Senn, 1994). A conspicuous imbalance is a diagnostic signal, not automatic invalidation. It should trigger an audit of the assignment code and its log, a check that no post-assignment filtering occurred, and a look at whether the imbalance appears across many variables or only one — a single covariate out of line with all the others, especially one related to geography or signup date, is the fingerprint of a randomizer keyed to something structured rather than random. If the randomization record is intact, chance imbalance does not remove the causal basis of the

intention-to-treat comparison, and the correct optional response is a covariate-adjusted analysis declared in advance for precision, not a re-randomization, which would itself introduce selection.

The third instrument is the one that grades the pipeline rather than the sample.

DEFINITION

A/A Test

An A/A test is an experiment in which both arms receive the identical experience, run through the full assignment, delivery, and analysis pipeline. Because there is no treatment, the true effect is zero, and the test therefore grades the apparatus rather than the intervention: assignment rates should match the intended split, pre-assignment characteristics should balance, and the measured difference should be statistically significant at approximately the nominal rate rather than more often. Because the nominal rate is not zero, a single significant A/A result is an occasion for investigation rather than a verdict; repeated or systematic failures are strong evidence of a defect in randomization, delivery, or measurement.

Source: Adapted from Kohavi et al. (2013).

The distinction the definition insists on matters because this chapter's lab does something narrower than a genuine A/A test, and the draft conflated the two. A true A/A test sends both arms through the production randomizer, the production message delivery, the production tracking, and the production campaign configuration, and it must be run before or alongside real experiments to certify that machinery. What Lab 11.1 Part B does is take completed control-arm outcomes and divide them at random into two sham groups. That procedure checks the analysis path — the metric definition, the test statistic, the code — and it checks the false-positive behavior of the test under a known-zero effect. It checks nothing whatever about the randomizer, the delivery, the tracking, or the assignment-to-exposure link, because none of those touched the sham split. This guide therefore calls the lab procedure a sham-split analysis-path check, reserves "A/A test" for the end-to-end version, and asks students to state which of the two they have run whenever they report one.

CONCEPT

Declare It or Discover It, Never Both

Every element of an experiment is either declared before the outcomes are visible or discovered after them, and the same number means different things depending on which. A 1.8-point lift on the metric named in Table 11.3 is a tested prediction. A 1.8-point lift on the metric selected from among nine after the results arrived is a description of the largest of nine numbers, and its interval — computed as if it were the only one — is wrong by an amount nobody can calculate.

The rule this guide enforces is not that discovery is forbidden; exploration after an experiment is legitimate, valuable, and how the next hypothesis gets built. The rule is that discovered findings are labeled as discovered, reported in the exploratory section of the deliverable, never given a p-value that pretends they were predicted, and never used as the basis for a decision until a subsequent experiment declares them in advance and tests them properly. Two words on a slide — "primary" and "exploratory" — carry the entire distinction, and an analyst who applies them honestly will occasionally lose an argument in the short run and will never have to retract a rollout.

Source: Course concept developed for this guide, informed by Nosek et al. (2018) and Simmons et al. (2011).

11.6 Measuring Lift

The design is now complete and the outcomes are in: the experiment ran, the outcome window closed, and the ab_test file contains one row per assigned customer with the arm, the pre-assignment characteristics, and the conversion outcome. This section computes the answer. It is the shortest conceptual section in the chapter, because the analysis of a well-designed experiment is deliberately simple — the credibility was purchased in Sections 11.4 and 11.5, and the arithmetic merely collects it.

Both arms are counted at the grain of assignment, and every randomized customer with an observed outcome is counted in the arm she was assigned to. Of the 5,000 customers assigned to control, 300 converted within the fourteen-day window, a rate of 6.00 percent. Of the 5,000 assigned to treatment, 390 converted, a rate of 7.80 percent. Everything the chapter needs follows from those four numbers, and the first decision is how to express the difference between them, because there are two conventions and they produce numbers that differ by a factor of seventeen.

DEFINITION

Absolute Lift and Relative Lift

Absolute lift is the difference between the treatment and control rates, expressed in the units of the metric — for a conversion rate, in percentage points. Relative lift is that same difference expressed as a proportion of the control rate, usually reported as a percentage change. Absolute lift is the quantity that scales to business volume, since multiplying it by the number of customers reached yields the incremental count directly. Relative lift expresses the effect against its own baseline, which can support comparison across settings when a multiplicative-effect assumption is plausible; transport across populations, periods, or doses remains a substantive assumption rather than a property of the statistic. Relative lift is systematically the larger and more persuasive of the two whenever the baseline rate is small, and reporting either without the other, or without the baseline rate, invites misreading.

Source: Adapted from Kohavi et al. (2020).

For the Drop Alert, the absolute lift is 7.80 minus 6.00, or 1.80 percentage points, and the relative lift is 1.80 divided by 6.00, or 30 percent. Both are correct. The deck that opens this chapter reported a relative number for the observational comparison — "more than 110 percent" — and the choice was not innocent, because a relative lift on a small base is the largest, most quotable version of any marketing result and the one least connected to money. The guide's standing rule from here forward: report the absolute lift first, in percentage points, with the control rate stated beside it, then the relative lift, then the translation into incremental conversions and dollars at the rollout's scale. An analyst who reports "30 percent lift" and stops has told a stakeholder almost nothing, since 30 percent of an unstated base is not a quantity.

The definition's middle sentence corrects a claim the draft of this chapter made too strongly. It is often said that relative lift is the quantity that transfers across populations with different baseline rates, and the sentence has the form of a statistical fact while actually being an empirical hope. A relative lift transfers only if the treatment's effect really is multiplicative in the baseline — if a program that raises a 6 percent rate by 30 percent would also raise a 2 percent rate by 30 percent. Sometimes that holds approximately. Sometimes the effect is closer to additive, and sometimes it is neither, because the mechanism that makes a treatment work in one population is absent in another. The honest position is that relative lift can support cross-setting comparison under a stated multiplicative assumption, and that the assumption belongs in the deliverable's limitations rather than in the definition of the statistic.

The estimate now needs its uncertainty, and this is where Section 7.10's evidence reading extends to a new kind of statistic rather than being re-derived. The logic is identical: 1.80 points is an estimate from one draw of ten thousand customers, another draw would have produced a somewhat different number, and the standard error measures that wobble. What differs is only the formula, because the outcome is a proportion rather than a continuous spend figure: the standard error of a difference in two proportions is built from each arm's rate and size, and for StyleCraft's numbers it works out to 0.5066 percentage points. The 95 percent confidence

interval on the difference is the estimate plus or minus roughly two standard errors, which gives a range from 0.81 to 2.79 percentage points. The corresponding test of the null hypothesis that the two population rates are equal returns a test statistic of 3.55 and a two-sided p-value of 0.0004 — the data are hard to reconcile with "the Drop Alert does nothing" — and the guide's position on that p-value is exactly the position Section 7.10 took, adopted from the American Statistical Association's consensus statement: it is a measure of surprise under the null, not the probability that the finding is true, not a measure of the effect's size, and not by itself a warrant for a decision (Wasserstein & Lazar, 2016).

What carries the decision is the interval, and reading it well is the section's real skill. The interval says that lifts as small as 0.81 points and as large as 2.79 points are compatible with what the experiment observed. That is a range of more than three to one in business terms, which is worth sitting with: a well-powered experiment on ten thousand customers, correctly designed and cleanly executed, still leaves the effect known only to within a factor of three. Two habits follow. The first is to check where the true value sits — and here the chapter's verification theme pays off, because the designed lift of 1.8 points that the generator planted lies comfortably inside the interval, near its center, which is what a correct analysis of a correct experiment should produce and what the lab requires students to confirm. The second is to translate both ends of the interval into the decision's units before drawing any conclusion, because a range that is wide in percentage points may be irrelevant in dollars or decisive in dollars depending entirely on the program's economics. That translation is Section 11.7.

Two technical notes belong here and then get out of the way, and the second is a correction the draft needed. First, the interval above uses each arm's own rate to estimate the standard error, while the significance test conventionally pools the two arms under the null; for StyleCraft's numbers the two give 0.5066 and 0.5069 percentage points respectively, a difference visible only in the fourth decimal, and the guide does not ask students to choose between them beyond knowing that the software's default is one of the two and that both are reported in the lab. Second, an interval for the relative lift may not be obtained by dividing the absolute interval's endpoints by the observed control rate. That shortcut treats the control rate as a known constant when it is itself an estimate with its own sampling error, and it therefore produces an interval that is too narrow and not centered where a ratio interval should be. For StyleCraft's numbers the shortcut returns roughly 13 percent to 47 percent, while a properly constructed 95 percent interval for the ratio of the two proportions — computed on the log scale and exponentiated, which is the standard construction — runs from a risk ratio of 1.12 to 1.50, or a relative lift of roughly 12 percent to 50 percent. The difference is not large here, and the discipline is the point: a confidence interval for a ratio is computed as a ratio. Code 11.9 computes both and prints them side by side so students can see how far the shortcut lands from the honest answer. The absolute lift remains the primary result throughout, and the relative interval is reported second, precisely because a stakeholder who has been told "30 percent" will be surprised to learn the evidence is consistent with 12.

11.7 Statistical Significance, Practical Significance, and the Break-Even Line

Section 11.6 produced an interval and a p-value. This section converts them into a decision, and the conversion turns on a distinction the guide has raised before in weaker forms and now makes operational: the difference between a result the data can distinguish from zero and a result the business should act on. The two questions are unrelated, they are answered by different quantities, and confusing them is the most common analytical error in commercial experimentation.

Statistical significance answers a question about evidence: given the sample, is the observed difference larger than the wobble that random assignment alone would produce? Practical significance answers a question about consequences: is a difference of this size worth what it costs to obtain? The first is settled by the interval's distance from zero; the second is settled by the interval's position relative to a line the analyst must compute from the program's economics before the experiment reports anything. Both failure directions occur regularly. A very large test — hundreds of thousands of customers, as platform experiments routinely run — will return significant results for effects of one-hundredth of a percentage point, differences that are unambiguously real and completely worthless. A small test on a genuinely valuable treatment will return a wide interval straddling zero, and the correct reading is not "the treatment does nothing" but "this experiment could not tell," which is a statement about the experiment, per Section 11.5's power discussion.

The instrument that makes practical significance concrete is the break-even lift, and it cannot be computed at all until the treatment's dose is fixed.

CONCEPT

The Break-Even Lift, and Why the Dose Comes First

Every marketing program has a cost per customer reached and a contribution per incremental conversion, and their ratio is the lift at which the program exactly pays for itself. The ratio is undefined until someone states how many times the program touches a customer, because the cost per customer reached is a cost per exposure multiplied by a number of exposures. The Drop Alert costs twelve cents per alert; Table 11.3 declares a dose of one alert per treated customer, delivered for the single drop in the customer's assignment week, with conversion measured over the fourteen days that follow. Total cost per assigned customer is therefore $0.12, StyleCraft's average order contributes about $28 after product cost, and the break-even lift is $0.12 ÷ $28 = 0.0043, or 0.43 percentage points.

Change the dose and the line moves. Had the test delivered two weekly alerts across its fourteen-day window, the cost per assigned customer would have been $0.24, the break-even lift would have been 0.86 points, and the same measured interval would no longer clear it at its lower end — the recommendation would change, on identical evidence, because the program being priced would be a different program. This is why the dose belongs in the registration beside the metric, and why the rollout's dose must be stated and compared: an experiment run at one alert per drop licenses a rollout at one alert per drop, and a rollout that raises the frequency is a different intervention whose economics and whose response must both be re-derived.

The comparison that follows is made against the whole interval, not the point estimate. If the interval's lower bound exceeds the break-even line, the program pays even in the pessimistic world the data still permits, and the decision is robust to the uncertainty rather than dependent on the estimate being exactly right. If the line falls inside the interval, the honest report is that the evidence does not settle the question, and the choice belongs to the decision-maker under the asymmetric-cost reasoning of Section 2.7 — which is a different sentence, and a more defensible one, than a recommendation dressed in a p-value.

Source: Course concept developed for this guide, informed by Kohavi et al. (2020).

Run the comparison for the Drop Alert and the decision becomes unusually clean. The break-even lift is 0.43 points. The experiment's interval runs from 0.81 to 2.79 points. The entire interval sits above the line, which means that every value of the effect the data considers plausible is a value at which the program pays for itself — and the recommendation therefore does not depend on the point estimate being right. Priced per thousand customers enrolled at the tested dose, the arithmetic reads: at the point estimate of 1.80 points, eighteen incremental conversions per thousand, contributing $504 against a program cost of $120, a net of $384; at the interval's pessimistic edge of 0.81 points, roughly eight incremental conversions, contributing $226 against the same $120, a net of $106; at the optimistic edge of 2.79 points, about twenty-eight conversions, $782 against $120, a net of $662.

That arithmetic is exact and its inputs are not, which the deliverable must say in the same breath. Below the break-even lift the program loses money and above it the program makes money under a specific set of assumptions, and the phrase "at any scale" that the draft of this

chapter attached to those clauses is not one the arithmetic supports. The calculation assumes that treatment cost is linear in customers reached, that contribution per incremental conversion is stable, that there is no fixed implementation cost to amortize, that no capacity or inventory constraint binds at rollout volume, that order value and margin do not change, that unsubscribe and fatigue costs are zero, and that response does not diminish as the program scales. Every one of those assumptions is plausible at the tested dose and volume, and several of them are known to weaken at rollout. The $28 contribution figure deserves particular scrutiny, because it is an average across observed orders, and the orders an incremental converter places need not carry the same category mix or margin as the average order. That is exactly why revenue per assigned customer is a declared secondary outcome in Table 11.3 rather than an afterthought, and why the deliverable presents the break-even comparison with a sensitivity band — the line recomputed across a plausible range of contribution and cost — rather than as a single number. Code 11.12 prints that band.

The honest sentence for the CMO is therefore not "the Drop Alert lifts conversion 30 percent" but something closer to: at a dose of one alert per drop, the program is profitable across the full range of effects the experiment supports, with a most likely value near $384 per thousand customers per drop window and a plausible range from roughly $106 to $662 under the stated cost and contribution assumptions — and the decision to expand does not hinge on which end proves true.

That is a strong result, and this is the moment to notice what it did not license. Three restrictions travel with it, all of which Section 11.13 will develop into the deliverable's language and none of which the p-value or the interval expresses.

The first is population. The experiment ran among SMS-opted-in customers with accounts at least thirty days old, which is a population selected for exactly the engagement that predicts response. The CMO's proposed rollout reaches the full SMS-eligible base, which is broader. Nothing in the experiment estimates the effect among customers outside its eligible population, and the reasonable prior — supported by the observational table of Section 11.3, where the adjusted difference happened to be identical in both engagement strata but need not have been — is that the effect on less engaged customers could be smaller, larger, or differently signed. The estimate is about the tested population, full stop, and the extrapolation to a broader one is an assumption the deliverable must name rather than a finding the experiment supports.

The second is dose and duration, and Section 11.7's break-even discussion has already shown how much of the business case rides on it. The experiment measured one alert and one fourteen-day outcome window in a non-holiday period. The rollout would run for a quarter, through the holiday calendar, at an alert frequency the CRM team has proposed increasing. Effects measured at one exposure routinely fail to survive a frequency increase, because unsubscribe and fatigue are nonlinear, and effects measured over two weeks routinely fail to persist over three months, for reasons Section 11.8's novelty discussion makes precise. Doubling the dose doubles the cost with certainty and does not double the response with anything like the same certainty, which is a specific and testable claim the rollout plan must schedule a measurement for.

The third is the counterfactual on the other side of the budget. The CMO's proposal moves four hundred thousand dollars out of paid social. The experiment tells StyleCraft what the Drop Alert is worth; it says nothing whatever about what the paid social spend is worth, and comparing a measured program against an unmeasured one systematically favors the measured one — which, as Section 11.14 will show, is one of the most consequential biases in modern marketing budgeting. The correct deliverable states plainly that the reallocation decision requires an incrementality estimate for paid social that StyleCraft does not currently have, and recommends the experiment that would produce it. This is not evasion. It is the analyst refusing to let a well-measured 1.8 points beat an unmeasured number by default.

11.8 How Tests Go Wrong

An experiment's credibility is established by design and destroyed by execution, and the destruction is almost always inadvertent. This section catalogs the five failures that account for most of the wreckage in commercial testing, each with the mechanism that produces it, the symptom it leaves in a report, and the defense. The first two are analytical, committed by the person reading the results; the third is operational, committed by the team running the test; the fourth and fifth are substantive, committed by reality.

The first is peeking, and it is the most common and the most misunderstood.

DEFINITION

Peeking

Peeking is the practice of examining an experiment's results before the pre-declared sample size or duration has been reached and allowing what is seen to influence when the experiment stops or how it is analyzed. Because the estimate fluctuates as data accumulate and each look offers a fresh opportunity to observe a fluctuation that crosses the significance threshold, the probability of declaring a difference where none exists rises with the number of looks — far above the nominal level the test's p-value reports. The resulting false positives are not detectable in the final output, since the test statistic computed at the moment of stopping looks entirely ordinary.

Source: Adapted from Johari et al. (2017).

The mechanism deserves to be felt rather than merely stated, which is why Lab 11.2 makes students manufacture a false winner from StyleCraft's own control arm. Imagine a test in which the treatment does nothing at all. On any single look at the accumulated data, there is a 5 percent chance the difference crosses the conventional threshold by luck. An analyst who looks once and stops has a 5 percent false-positive rate, as advertised. An analyst who checks the dashboard each morning for two weeks and stops the moment the result "reaches significance" has taken fourteen opportunities to catch a 5 percent event, and while the looks are correlated — each day's data contains the previous day's, so the arithmetic of fourteen independent trials, which would suggest 51 percent, badly overstates the damage — the inflation remains severe. On StyleCraft's own control arm at fourteen evenly spaced looks, the probability of seeing a significant

difference in either direction at some point rises to roughly 22 percent, more than four times the nominal rate, and Lab 11.2 has students measure it.

The organizational version is worse than the statistical one, because the stopping is asymmetric. Nobody stops a test early because the treatment is losing; tests are stopped early because the treatment is winning, which converts peeking from a symmetric random error into a machine that manufactures winners. That asymmetry is worth measuring separately rather than folding into the two-sided figure, and the lab does so: the probability of at some point seeing a significant difference with the treatment arm ahead — the policy an actual analyst follows — runs near 11 percent on this design, roughly one test in nine. The two-sided 22 percent and the favorable-only 11 percent answer different questions, and reporting one while describing the other is a small dishonesty the draft of this chapter committed and Code 11.10 now separates.

The defense declared in Table 11.3 is the simplest available: fix the sample size and the duration in advance, look once when the outcome window closes, and treat the dashboard between launch and close as an operational monitor for delivery failures rather than a scoreboard. Sequential methods that permit valid continuous monitoring exist and are standard at platform scale (Johari et al., 2017), but they must be declared in advance too — a test analyzed with an ordinary two-proportion comparison after ad hoc looking is not rescued by mentioning that such methods exist.

The second failure is peeking's twin, committed across metrics rather than across time.

DEFINITION

Multiple Comparisons

The multiple comparisons problem is the inflation of false-positive risk that occurs when many hypotheses are tested and any significant result among them is treated as though it had been the only test performed. With independent tests at a 5 percent threshold, the chance of at least one spurious significant result is roughly 64 percent after twenty tests and rises toward certainty thereafter, so a "winning" metric or segment selected from a large set carries far weaker evidence than its p-value suggests. Remedies include declaring a single primary outcome in advance, adjusting thresholds for the number of comparisons, and treating unplanned findings as hypotheses for a future experiment rather than as results.

Source: Adapted from Simmons et al. (2011) and Kohavi et al. (2020).

In marketing the multiplicity is rarely deliberate and almost always large. A single experiment report may carry conversion, revenue per customer, average order value, order count, app opens, and margin — six metrics — each cut by four segments, two channels, and three loyalty tiers, which is not six tests but well over a hundred. Somewhere in that grid, something is significant. The symptom in a report is unmistakable once you know to look for it: a headline that describes a subgroup rather than the population ("the Drop Alert lifted conversion 6.2 points among Gen Z suburban app users"), attached to an interval computed as if that subgroup had been the plan all along. The defense is Table 11.3's primary-metric row and Section 11.5's labeling discipline. One

primary metric decides. Everything else is exploratory, is labeled exploratory in the deliverable, and generates the next preregistration rather than this quarter's rollout.

The third failure is operational, and it is the reason experiment platforms have change logs. Mid-test changes are alterations to the treatment, the control, the eligible population, or the measurement while the experiment is running: a creative refresh applied to the treatment arm in week two, a bug fix that changes what the control experiences, an eligibility filter tightened after launch, a tracking change that alters how conversions are counted. Each of them silently redefines what is being compared, so that the final result estimates the effect of neither the original treatment nor the revised one but of an undocumented blend whose composition depends on when customers happened to be assigned. The symptom is a result that changes character across the test window without an obvious substantive reason, and the defense is a change log plus a hard rule that any change to the treatment restarts the experiment. StyleCraft's registration handles the common marketing version of this failure explicitly: the eligible population and the message creative were fixed before launch, and the CRM team's ordinary weekly optimization was suspended for both arms during the test window.

The fourth failure is substantive, and it is the one most likely to cost a marketing organization real money after a correct experiment.

DEFINITION

Novelty Effect

A novelty effect is a response to a treatment that arises from its unfamiliarity rather than from its enduring value, producing an initial effect that decays as the audience habituates. Its mirror image, sometimes called a primacy effect, is an initial underperformance caused by disruption of established habits, which fades as users adapt. Because both concern how response changes with repeated exposure or with time since first exposure, detecting either requires outcomes observed across multiple exposures or multiple periods for the same randomized units — not a comparison of units first exposed at different dates.

Source: Adapted from Kohavi et al. (2020).

The Drop Alert is an unusually good candidate for the effect. A twelve-hour early-access window is interesting the first time a customer receives it and steadily less interesting the twentieth time, particularly if the customer discovers that the styles she wanted were still available at the general announcement anyway. An experiment measuring one alert and one fourteen-day window is therefore measuring the alert at its most novel, and the honest expectation is that a quarter-long rollout at higher frequency returns something smaller than 1.8 points.

What the closing sentence of the definition rules out is the diagnostic the draft of this chapter proposed, and the correction is worth stating plainly because the flawed version is everywhere in commercial practice. Splitting an experiment's customers into those assigned in the first half of the assignment calendar and those assigned in the second, then comparing the estimated lift in each, does not test for novelty. It compares two different cohorts of customers, assigned on

different dates, exposed to different drops, under different promotional and calendar conditions. Any difference between the halves can be cohort mix, drop attractiveness, promotion timing, weather, or — given roughly 2,500 customers per arm per half — plain noise. What the comparison can honestly be called is a temporal-stability diagnostic: a crude check for whether the estimate wobbles alarmingly across the window, useful only for flagging a difference dramatic enough to demand investigation and incapable of certifying that no decay exists. This guide labels it that way, and Code 11.13 prints it under that name.

Estimating a novelty effect properly requires a design the certified ab_test file does not support, and saying so is more useful than pretending otherwise. The file carries one row per customer with a single binary outcome over a single fourteen-day window, which contains no repeated-exposure information at all. The design that would answer the question retains one randomized assignment across several consecutive drops, records the outcome after the first exposure, the second, the third, and so on, estimates a treatment-by-period interaction, and schedules a longer-run follow-up after the program has stopped being new. That design is what the rollout plan must commission, and Exercise 11.4 asks students to specify it.

The fifth failure is the design-time contamination question of Section 11.4, arriving as an execution problem.

DEFINITION

Contamination

Contamination — also called interference or spillover — occurs when a unit's outcome is influenced by the treatment assigned to other units, violating the assumption that each unit's outcome depends only on its own assignment. Common marketing pathways include treated customers sharing offers with untreated ones, treated and untreated customers competing for the same limited inventory, household or device sharing that mixes arms, and market-level effects such as auction prices that respond to a treated group's behavior. Contamination biases the measured difference — usually toward understating the treatment's true effect when the control arm partially receives it, and toward overstating it when the treated arm's activity depletes something the control arm needed — and it is addressed by choosing a larger unit of randomization rather than by any adjustment made after the fact.

Source: Adapted from Kohavi et al. (2020).

Marketing's canonical contamination case is the promotion code that leaks: a treatment-arm offer posted to a deals forum and redeemed by hundreds of control-arm customers, which pulls the two arms toward each other and shrinks the measured effect toward zero. The symptom is a suspiciously small effect combined with control-arm behavior that should not have been possible, and the defense at design time is either a larger unit — household, store, market — or a treatment that cannot travel, such as a personalized link. StyleCraft's version, noted in Section 11.4, runs the other way: early access to limited inventory could deplete what the control arm finds later. Its expected direction was recorded in the registration precisely so the estimate could be read with the bias's sign known rather than debated afterward.

Table 11.5 collects the five failures with their mechanisms, their symptoms, and their defenses. It is written to be used rather than read: the symptom column is what an analyst actually encounters — somebody else's test report, a vendor's case study, a slide from another team — and Exercise 11.7 is nothing but this column, applied to six sentences.

Table 11.5

The five standing ways marketing tests go wrong

Failure	Mechanism	Symptom in the report	Defense
Peeking	Repeated looks at accumulating data, with asymmetric early stopping on favorable results	A test that stopped at an unplanned moment; a duration that does not match the plan; a winner that does not reproduce	Fixed sample size and duration; a single analysis at close; declared sequential methods if monitoring is required
Multiple comparisons	Many metrics and segments tested, the largest reported as if planned	A headline about a subgroup; a p-value attached to a finding nobody predicted	One declared primary metric; everything else labeled exploratory and re-tested before it is acted on
Mid-test changes	Treatment, control, population, or measurement altered while the test runs	Effects that shift character mid-window; a definition that changed after launch	Change log; freeze on optimization during the test; any treatment change restarts the experiment
Novelty effect	Response driven by unfamiliarity, decaying with repeated exposure	A strong single-exposure effect that fades after rollout at higher frequency	A design that follows one randomized assignment across several exposures; a treatment-by-period estimate; a planned re-measurement after habituation
Contamination	Treatment reaches or affects control units	An implausibly small effect; control-arm behavior the design should have prevented	Larger unit of randomization; non-transferable treatments; expected direction declared in advance

11.9 When You Cannot Randomize: Quasi-Experimental Intuition

Everything so far has assumed the analyst can decide who is treated. Frequently she cannot. A store's assortment change applies to the store; a regional out-of-home campaign applies to the

metro; a pricing policy applies to the catalog; a competitor's entry applies to whoever it applies to. This section sketches, carefully and at concept level, what can be salvaged when randomization is unavailable — and it is bounded deliberately, because the designs below are the entry point to a large technical literature and this guide teaches only enough to read such evidence critically and to recognize when a specialist is required.

Begin by disposing of the weakest design, which is also the most common in marketing decks. A simple before-and-after comparison — revenue in the eight weeks after the campaign against the eight weeks before — is a causal claim with no comparison group at all, and its implicit counterfactual is that nothing else would have changed. In a business with the seasonality of Section 10.3, an expansion producing structural breaks per Section 10.4, and a promotional calendar that moves demand around inside the quarter, that counterfactual is not merely uncertain but usually false. A pre/post comparison is worth computing and worth never reporting alone.

The repair is to add a comparison group that experienced the same period without the treatment: other stores, other metros, other customer cohorts. The comparison group supplies an estimate of what would have happened anyway, and the treated group's change is judged against it rather than against zero. This is difference-in-differences, and it is best learned as arithmetic before it is learned as a method.

DEFINITION

Difference-in-Differences

Difference-in-differences estimates a treatment effect by comparing the change in outcomes over time in a treated group with the change over the same period in an untreated comparison group, and taking the difference between the two changes. The comparison group's change stands in for what the treated group's change would have been absent the treatment, so any level difference between the groups that is stable over time is removed by the design. The estimate's validity rests on the parallel trends assumption: that in the absence of the treatment, the two groups' outcomes would have moved in the same direction by the same amount. This assumption concerns an unobserved counterfactual and therefore cannot be verified — only supported, most commonly by showing that the groups' outcomes moved together over multiple periods before the treatment began.

Source: Adapted from Card and Krueger (1994) and Angrist and Pischke (2015).

Suppose StyleCraft's suburban New York metro receives a regional out-of-home campaign that cannot be aimed at individuals, and the analyst designates the Boston and Chicago flagship markets as the comparison. Revenue per active customer in the treated metro runs $82 in the eight weeks before and $95 in the eight weeks after, a change of $13. In the comparison markets it runs $78 and $86, a change of $8. The difference-in-differences estimate is $13 minus $8, or $5 per active customer — and the $8 is the whole point of the exercise, since a pre/post report would have credited the campaign with $13 and quietly attributed the season, the assortment refresh, and the national brand momentum to a billboard.

What the design assumes is where the caution lives. Parallel trends is a claim about a world that did not happen, and it fails routinely in exactly the situations marketing cares about most: the metro chosen for the campaign was often chosen because something was already happening there, which is selection bias operating on markets rather than customers, and the comparison markets differ in store maturity, competitive intensity, and customer mix in ways that may make their trends diverge for reasons unrelated to any treatment. The evidence that supports the assumption — never proves it — is the pre-period, and the standard practice is to plot several periods of both groups before the treatment and confirm they move together. If they do not, the design is not usable and no amount of statistical machinery repairs it. Athey and Imbens (2017) survey the modern methods that extend this family, and their existence should be read as a signal rather than a menu: these designs are a specialty, their assumptions are contestable, and a marketing analyst's job is usually to recognize that a quasi-experimental claim is being made, ask what the comparison group is and whether the pre-trends were shown, and know when to bring in help.

The reason the section is bounded so tightly is worth stating plainly, because the temptation runs the other way. Quasi-experimental designs are attractive precisely because they promise causal language without the operational cost of an experiment, and that promise is what makes them dangerous in a marketing organization. Their conclusions rest on assumptions that cannot be checked, held by analysts who are motivated to believe them, about programs their own colleagues designed. The guide's working position: prefer an experiment whenever one is possible, use these designs when it is not, state the assumption in the deliverable in the same font as the result, and treat the resulting evidence as occupying the rung below a well-executed experiment on the ladder that Section 11.10 now builds.

11.10 The Evidence Hierarchy

This chapter began by distinguishing prediction from causation, and it ends its conceptual arc by making the distinction usable. The evidence hierarchy is a ranking of the kinds of support a marketing claim can have, and its purpose is not to rate methods but to constrain vocabulary: each rung licenses a specific set of verbs, and using a higher rung's verbs on a lower rung's evidence is the single most common form of analytical overclaiming, as Section 7.15 argued at length.

DEFINITION

Evidence Hierarchy

The evidence hierarchy is an ordering of sources of evidence by the strength of the causal claim each can support: anecdote and observation at the base, then measured association, then association estimated with statistical controls or predictive validation, then quasi-experimental designs resting on stated assumptions, and finally randomized experiments at the top. Each rung narrows the set of rival explanations that remain live. The hierarchy describes the warrant available from competently executed designs; design labels alone do not determine evidential quality. It is a guide to what may be asserted, not to what is useful — evidence at every rung has legitimate roles, including generating the hypotheses that higher rungs test — and its practical function is to hold the language of a conclusion to the strength of the design that produced it.

Source: Adapted from Angrist and Pischke (2015) and Kohavi et al. (2020).

That qualification in the middle of the definition is not diplomatic hedging; it is the difference between a hierarchy and a badge. A randomized experiment with severe outcome attrition, widespread noncompliance, uncontrolled interference between arms, a primary metric switched after the results arrived, or a stopping decision made on the fourteenth peek can easily provide weaker evidence than a carefully executed difference-in-differences whose pre-trends were shown and whose assumption was stated. The rung a study occupies is set by its design; the warrant it actually delivers is set by its design and its execution together. This is why Table 11.4's assignment-flow record and Section 11.5's integrity checks are not paperwork attached to the top rung — they are what entitles a study to claim it.

Table 11.6 places the guide's own output on the ladder. Reading it is a useful shock, because the great majority of what Part II produced sits on the third rung, and the chapters that produced it were the most technically demanding in the book.

Table 11.6

The evidence hierarchy, with this guide's own output placed on it

Rung	What it is	Where it appears in this guide	The strongest sentence it licenses
1. Anecdote	A case, a quotation, a memorable single observation	The merchant's conviction (Section 10.1); "customers keep telling us…"	"Here is something that happened, which may be worth investigating."
2. Description and association	Rates, cross-tabs, correlations measured in observed data	Chapters 5 and 6; the correlations of Section 7.3; the 5.30-point difference of Section 11.1	"These things occur together in our data, by this much."

Rung	What it is	Where it appears in this guide	The strongest sentence it licenses
3. Model-based association and prediction	Multivariable estimates with controls; out-of-sample predictive validation; forecasts	Chapters 7, 8, 9, and 10; the adjusted 1.80-point difference of Section 11.3	"Holding these measured factors constant, X accompanies Y by this much," or "this model predicts Y this accurately on data it has not seen."
4. Quasi-experiment	Pre/post with a comparison group; difference-in-differences and its relatives	Section 11.9	"Under the stated parallel-trends assumption, the treatment appears to have changed the outcome by this much."
5. Randomized experiment, competently executed	Random assignment to treatment and control, analyzed as assigned, with assignment integrity documented	Sections 11.4 through 11.7; the ab_test deliverable	"Among the tested population, in the tested window, at the tested dose, the treatment caused a change of this much, within this interval."

Three readings of the table matter more than the table itself. The first is that the top rung is narrow. Even a clean experiment licenses a sentence that is heavily qualified by population, window, and dose, and the qualifications are not throat-clearing — they are the boundary of what was tested, and Section 11.13 will show what happens when they are dropped in the retelling. The second is that the lower rungs are not the losers. Chapter 6's segments could not have been discovered by an experiment; Chapter 9's churn model does a job no experiment does, which is ranking individuals for action; Chapter 10's forecast answers a question about the future that no experiment addresses. The hierarchy ranks causal warrant, not usefulness, and an organization that ran only experiments would be blind. The third and most practical is that the hierarchy is a vocabulary constraint that can be checked in a draft. Verbs like *drives*, *causes*, *lifts*, *generates*, and *produces* are rung-five verbs; *is associated with*, *accompanies*, *predicts*, and *is higher among* are rung-two and rung-three verbs. An analyst editing a deck for these words, once, before it ships, will catch nearly every overclaim she is about to make.

CONCEPT

The Ladder Is a Vocabulary Constraint

The practical discipline this chapter leaves behind outlives its arithmetic, and it is small enough to apply in a hallway. For any sentence about to be spoken to a decision-maker, ask which rung produced it, then check whether the verb belongs to that rung. "The loyalty program drives a 22 percent higher repeat rate" is a rung-five sentence resting on rung-two evidence, and the fix is not to hedge it into meaninglessness but to say the true thing: "Loyalty members repeat 22 percent more often than non-members; we have not tested whether enrolling a customer causes that, and the customers who enroll are already our most frequent."

Stakeholders do not resent the second sentence. They resent discovering, two quarters into a program built on the first one, that nobody had ever checked. The analyst who speaks in rung-appropriate verbs from the beginning is the one whose rung-five sentences get believed when she finally has one.

Source: Course concept developed for this guide, informed by Angrist and Pischke (2015).

11.11 AI as an Experimentation Assistant

The division of labor the guide has refined since Chapter 7 — the assistant drafts mechanics, the analyst audits meaning — arrives in this chapter at a task where the assistant's competence and the task's requirements are almost perfectly misaligned, and the misalignment is worth naming before the failure modes are listed. An experiment's arithmetic is trivial and its epistemology is everything. Assistants are excellent at arithmetic and structurally incapable of the epistemology, because the epistemology lives in facts that are nowhere in the data: whether the assignment was random, whether the metric was declared before the outcomes were seen, whether this is the first look or the fourteenth, whether the subgroup in the prompt was chosen before or after somebody noticed it was winning. Hand an assistant two columns and ask whether the difference is significant and it will answer — correctly, fluently, with an interval and a p-value and a sentence about the result being statistically significant — and it will answer in exactly the same way whether the data came from Section 11.4's randomized design or from Section 11.1's targeted send, whether the test was declared or discovered, and whether it is being run for the first time or the fortieth. Nothing in the numbers reveals the difference, and nothing in the assistant's training makes it ask.

The experimentation-specific failure modes follow from that gap, and they are named here for the audit. The observational file analyzed as a trial: A/B machinery applied to data in which somebody chose who was treated, producing a confident causal sentence about the selection bias of Section 11.3 — this is the highest-frequency and highest-cost failure, and it is invited by the request itself whenever the prompt says "campaign" instead of "experiment." The uncomplaining test on peeked data: the assistant computes significance for whatever window it is handed and will not ask when the test started, how many times it has been examined, or whether the analyst is stopping now because of what she just saw. The invented design: asked to design a test, the

assistant fills every gap the prompt left — inventing a metric definition, choosing a duration, assuming a unit of randomization, assuming a treatment dose, sometimes assuming an eligibility rule — and does so silently and plausibly, so that the resulting registration reads like a plan the analyst made. The per-protocol drift: asked to analyze an experiment with delivery failures, the assistant will cheerfully drop the undelivered customers and never mention that it has just replaced the intention-to-treat comparison with an observational one. The segment harvest: asked to "check whether it worked for different customer groups," the assistant will happily test twenty subgroups and report the significant ones, executing Section 11.8's multiple comparisons problem at machine speed with no note that the p-values are meaningless as printed. The lift ambiguity: relative and absolute lift conflated or converted incorrectly, relative lift reported without the baseline, and a ratio interval produced by dividing a difference interval by the control rate, which is Section 11.6's shortcut committed by autocomplete. The naked winner: results reported as "treatment wins" with no interval, no practical-significance line, no stated dose, and no statement of the tested population — the assistant inherits the internet's preference for decisive conclusions. The grain slip: data randomized by customer but analyzed by order or session, inflating the apparent sample size and shrinking the interval to a width the design never earned. And the power omission: a null result narrated as "no effect" rather than "this test could not have detected an effect smaller than X," which is the Section 11.5 distinction the assistant will not make unless the prompt makes it.

Where the assistant genuinely helps, the list is long and worth using deliberately, because most of an experiment's labor is mechanical. Sample-size and power arithmetic is fully specifiable and the assistant is fast and accurate at it, including the sensitivity table of required sample size across a range of minimum detectable effects that makes the inverse-square relationship concrete for a planning meeting. Design checklists drafted against Table 11.3 catch omissions a human drafter skips. Balance-check, flow-table, and sham-split code is boilerplate. The two-proportion analysis, its two intervals, and the peeking simulation of Lab 11.2 are exactly the kind of well-specified mechanical work Section 8.9's frozen-rules discipline makes safe to delegate. Break-even arithmetic and the sensitivity band across contribution and cost are pure computation. And an assistant is genuinely useful as an adversary: asked to argue against a result — to list every rival explanation for an observed difference, every contamination path, every way the tested population or dose differs from the rollout's — it produces a better checklist than most analysts generate unaided, and the exercise costs nothing.

The governing instrument extends the audit lineage once more. Table 8.5's five points run unchanged — frame, leakage, split hygiene, baseline, error in decision units — with "baseline" reading, in this chapter, as the control arm and the break-even line. Table 11.7 adds the three checks experimentation makes necessary, and its first point is the one that has no analogue anywhere else in the guide, because it is a check on chronology rather than on code.

Table 11.7

The experimentation supplement to the five-point audit (run with Table 8.5)

Supplemental point	**The check**	**Fails when**
E1. Registration precedes analysis	A dated preregistration exists (Table 11.3), and the metric, arms, dose, unit, sample size, duration, decision rule, and stopping rule in the analysis match it exactly; deviations are listed as deviations	The metric or window in the analysis differs from the plan; no plan exists; the "primary" metric was chosen after the results were seen; the test stopped at an unplanned moment; the dose is nowhere stated
E2. Assignment integrity	Assignment was made by a stated chance mechanism independent of customer characteristics; the assignment-flow record of Table 11.4 is reported from randomization through outcome observation; the analysis is intention-to-treat; covariate balance is read as a diagnostic rather than a test; the analysis grain equals the assignment grain; contamination paths are named with expected direction	Anyone chose who was treated; undelivered or noncompliant customers were dropped from the primary comparison; a covariate imbalance is used to discard a valid randomization; sessions or orders are analyzed after customers were randomized
E3. Inference honesty	One primary metric analyzed once, with everything else labeled exploratory; absolute lift reported with the baseline rate beside it; a difference interval and a separately constructed ratio interval; the interval compared against a break-even line computed in advance from a stated dose, with its sensitivities; null results reported with their minimum detectable effect	Significance appears without an interval; a subgroup is the headline; relative lift is quoted alone or its interval is a rescaled difference interval; "no effect" is claimed by an underpowered test

AI IN PRACTICE

Preregister With the Assistant, Then Audit the Analysis Against the Registration

The two-prompt pattern of Sections 8.11, 9.11, and 10.11, restructured around this chapter's chronology, because here the order of the prompts is itself the control. Step one, before any data exists, is the registration prompt: "I am testing whether a twelve-hour early-access SMS lifts fourteen-day conversion among SMS-opted-in customers. Draft a preregistration covering every row of the following template — decision, primary metric with its four elements, arms including what control receives, treatment dose in number of exposures, assignment mechanism, unit of randomization, eligible population, sample size with the power calculation shown, duration, planned analysis, decision rule, stopping rule, secondary metrics labeled exploratory, and guardrails that are computable from the columns I have listed. Where the information I gave you is insufficient to fill a row, write UNSPECIFIED and ask me; do not fill it yourself." The final clause is the whole prompt, because the invented-design failure mode is defeated only by making the assistant's gaps visible.

Step two, the prediction, before the analysis runs: write down the lift you expect and the interval width you expect, computing the latter from the arm sizes and the baseline rate. The analyst who has done the power calculation already knows the interval will be roughly two points wide, and an interval that comes back at a tenth of that width means the grain slipped. Step three, the analysis prompt, which begins by pasting the registration back in: "Here is the registration. Analyze the attached file exactly as registered — intention-to-treat, no other metrics, no subgroups, one analysis. Report the assignment-flow counts, the control rate, the treatment rate, absolute lift in percentage points with a 95 percent interval on the difference, relative lift with a 95 percent interval on the ratio computed as a ratio, and the p-value. Do not interpret."

Step four, the audit: run Table 8.5's five points, then Table 11.7's three, line by line, and grade your prediction. Then the adversarial prompt, which is where the assistant earns its keep: "List every reason this estimate might not transfer to a chain-wide quarter-long rollout at two alerts per week instead of one, and every contamination path you can construct." Harvest that list into the deliverable's limitations. Only after all of it, the drafting prompt: "Write the rollout recommendation paragraph using only these numbers." Audit its verbs against Section 11.10's ladder — an experiment licenses *caused, among the tested population, at the tested dose, within this interval*; it never licenses *will* — file both exchanges per Appendix D, and sign the recommendation as the analyst of record.

Source: Course concept developed for this guide, informed by Kohavi et al. (2020) and Nosek et al. (2018).

11.12 Hands-On Application in Python and Google Colab

The preceding sections built the argument; this section runs it on StyleCraft's own files, in two labs that follow the chapter's chronology rather than its section order. Lab 11.1 establishes credibility before any effect is estimated: the observational trap reconstructed by hand at miniature scale so that the 5.30 points decomposes in front of the student, then the experiment's assignment audited for integrity with a data contract, an assignment-flow table, balance profiling, and a sham-split analysis-path check. Lab 11.2 sizes, estimates, and prices the effect: the power calculation that justified the sample, the registered intention-to-treat analysis with both of its

intervals, checked against the known true lift the dataset ships; the deliberate peeking simulation in which a false winner is manufactured from data containing no effect at all; the practical-significance comparison against the break-even line with its sensitivity band; a temporal-stability diagnostic that is carefully not a novelty test; and the round trip through the assistant that produces the rollout paragraph.

The labs use statsmodels throughout for the proportion tests, the interval constructions, and the power calculations (Seabold & Perktold, 2010); the three functions they depend on are documented by the statsmodels developers (2026a, 2026b). They use the certified ab_test file, whose columns follow the course data dictionary: one row per assigned customer, with customer_id, assign_date, variant (control or treatment), converted (1 if the customer placed at least one order within fourteen days of assignment), order_revenue for the same window, and the pre-assignment characteristics home_metro, generation, loyalty_tier, app_user, prior_orders_90d, and prior_spend_90d. Two conventions hold across every cell. Code cells are numbered continuously across both labs, so that the deliverable and the exercises can cite them. And every output is predicted before it is computed, per the standing rule. AI assistants may draft any code cell (Appendix C has templates; Appendix A covers Colab mechanics), and every exchange is documented per Appendix D.

11.12.1 Lab 11.1, Part A: The Observational Trap, Decomposed by Hand

The miniature this chapter needs is not a sample of rows but a table of four cells, because selection bias is a fact about composition and four numbers are enough to show it completely. The cells are the observational record of Table 11.1: sixteen thousand customers, crossed by whether they received the Drop Alert and by whether they were engaged at the time of the send. Before running anything, predict two quantities and write them down — the overall difference between the received and not-received groups, and the difference within each engagement stratum — and predict which of the two the CMO's deck reported.

Code 11.1. Aggregate the observational comparison

```
import pandas as pd

obs = pd.DataFrame({
    "engaged":     ["Engaged", "Engaged",
                    "Not engaged", "Not engaged"],
    "drop_alert":  ["Received", "Not received",
                    "Received", "Not received"],
    "customers":   [6000, 2000, 2000, 6000],
    "conversions": [708,  200,  96,   180],
})
obs["rate"] = obs["conversions"] / obs["customers"]

naive = (obs.groupby("drop_alert")[["customers", "conversions"]]
            .sum())
naive["rate"] = naive["conversions"] / naive["customers"]
print(naive)

recv = naive.loc["Received", "rate"]
notr = naive.loc["Not received", "rate"]
print(f"naive difference: {(recv - notr) * 100:.2f} pp")
print(f"relative:         {recv / notr - 1:+.1%}")
```

Expected output: the two-row aggregate table with customers, conversions, and rate for the received and not-received groups, then the naive difference in percentage points and the same difference expressed relatively.

Input is the crossed record; the transformation is a sum that discards the engagement dimension; the output is the number that started the argument, and the discarding is the point. Note that the four literal cells are the chapter's own Table 11.1, entered by hand so that every figure downstream can be checked with a calculator.

Code 11.2. Standardize by engagement composition

```
wide = obs.pivot(index="engaged", columns="drop_alert",
                 values="rate")
wide["within_diff_pp"] = ((wide["Received"]
                           - wide["Not received"]) * 100)
print(wide.round(4))

# What the treated group would have converted at with no
# campaign: its own engagement mix, at untreated rates.
# This is an ADJUSTED difference under the assumption that
# engagement is the only confounder -- not a causal effect.
w = (obs.loc[obs["drop_alert"] == "Received"]
        .set_index("engaged")["customers"])
w = w / w.sum()
counterfactual = (w * wide["Not received"]).sum()

print(f"treated actual:         {recv:.2%}")
print(f"treated counterfactual: {counterfactual:.2%}")
print(f"adjusted difference (pp): "
      f"{(recv - counterfactual) * 100:.2f}")
print(f"composition component (pp): "
      f"{(counterfactual - notr) * 100:.2f}")
```

Expected output: the stratified table with each arm's rate side by side and the within-stratum difference in percentage points, then the treated group's actual rate, its constructed counterfactual rate, and the split of the naive difference into an adjusted difference and a composition component.

The second cell does the work of Section 11.3 in a dozen lines. The pivot puts the two rates side by side within each engagement stratum; the weights w are the treated group's own engagement mix; and applying that mix to the untreated rates constructs the rate the treated group would have shown had the campaign never existed. Read the print labels carefully, because they are the correction this chapter's draft needed: the cell prints an adjusted difference and a composition component, not an effect and a bias. Calling the first quantity the campaign's effect would smuggle in four assumptions the record cannot check, and the experiment of Lab 11.2 is what makes those assumptions unnecessary rather than merely plausible.

VERIFICATION CHECK

Before you run: compute all of it by hand. The stratum rates are 708 ÷ 6,000 = 11.80 percent, 200 ÷ 2,000 = 10.00 percent, 96 ÷ 2,000 = 4.80 percent, and 180 ÷ 6,000 = 3.00 percent, so the within-stratum differences are 1.80 points and 1.80 points — identical, which is the designed signature of a record in which the treatment did the same thing to both kinds of customer. The aggregates are 804 ÷ 8,000 = 10.05 percent and 380 ÷ 8,000 = 4.75 percent, a naive difference of 5.30 points and a relative lift of 111.6 percent. The counterfactual is 0.75 × 10.00 + 0.25 × 3.00 = 8.25 percent, so the adjusted difference is 10.05 − 8.25 = 1.80 points and the composition component is 8.25 − 4.75 = 3.50 points.

After you run: confirm that the two components sum exactly to the naive 5.30, because the additivity is what makes the decomposition a decomposition rather than an adjustment. Then verify the composition component a second way, from mix alone: the treated group was 75 percent engaged and the untreated group 25 percent, a 50-point difference in mix, and engaged customers converted 7.00 points higher untreated (10.00 against 3.00), so 0.50 × 7.00 = 3.50 points arrive from nothing but who was chosen. Now write the sentence the miniature has earned, in the finance partner's language rather than the analyst's: two-thirds of the campaign's apparent performance was the CRM team's targeting skill, correctly measured and incorrectly attributed. Then write the second sentence, which is the one this chapter cares about more: the remaining 1.80 points is an adjusted observational difference that equals the campaign's effect only if engagement is the sole confounder, measured without error, and the only thing distinguishing the two groups — and the only reason the adjustment was possible at all is that engagement happened to be recorded, which is a piece of luck the next confounder will not repeat.

Investigate if: your hand arithmetic disagrees with the printed figures anywhere. Every number in these two cells is reproducible with a calculator, so a mismatch is arithmetic rather than convention.

11.12.2 Lab 11.1, Part B: Assignment Integrity — the Data Contract, the Flow Record, Balance, and the Sham Split

The experiment's credibility rests entirely on the claim that a chance mechanism, and nothing else, decided who received the Drop Alert. This part audits that claim before a single effect is estimated, which is the order the deliverable will also follow: an experiment whose assignment cannot be defended does not get to report a lift. Four cells do the work, and the first of them is a contract rather than an analysis.

Code 11.3. Load and validate the experimental file

```
import numpy as np
import pandas as pd

ab = pd.read_csv("ab_test.csv", parse_dates=["assign_date"])

# THE DATA CONTRACT. Every assertion is a claim the analysis
# depends on. A failure stops the lab; it is not repaired
# silently, and it is not repaired by dropping rows.
required = ["customer_id", "assign_date", "variant",
            "converted"]
missing = [c for c in required if c not in ab.columns]
assert not missing, f"columns absent from the file: {missing}"
assert ab[required].notna().all().all(), (
    "a required field is null: trace the source before "
    "computing anything")
assert ab["customer_id"].is_unique, (
    "customer_id repeats: this may be a duplicated RECORD or "
    "a customer randomized twice. Trace which before acting")
assert set(ab["variant"]) == {"control", "treatment"}, (
    f"unexpected arm labels: {sorted(set(ab['variant']))}")
assert ab["converted"].isin([0, 1]).all(), (
    "converted is not a clean 0/1 flag")

print("rows:", len(ab), " columns:", len(ab.columns))
print("assignment window:",
      ab["assign_date"].min().date(), "to",
      ab["assign_date"].max().date())
print(ab["variant"].value_counts().to_string())
```

Expected output: the row and column counts, the first and last assignment dates, and the count of customers in each arm; five assertions pass silently.

Input is the certified file; the transformation is nothing but validation; the output is permission to proceed. Read the duplicate assertion's message rather than just its condition, because it is a correction to the draft of this chapter. A repeated customer_id does not establish that a customer was randomized twice. It may equally be one randomization written to the file twice — a union join instead of a unique key, per Chapter 4's join disciplines — and the two defects have different consequences and different repairs. The code therefore stops and instructs the analyst to trace the source, rather than asserting a causal interpretation of a data defect it cannot distinguish.

Code 11.4. Build the assignment-flow table

```
# The flow record of Table 11.4. Randomized customers stay in
# the arm they were assigned to (intention-to-treat); nothing
# here is a filter, and delivery is reported, never applied.
PLANNED = {"control": 5000, "treatment": 5000}

flow = pd.DataFrame(index=["randomized",
                           "primary outcome observed",
                           "included in ITT analysis"])
for arm in ("treatment", "control"):
    g = ab.loc[ab["variant"] == arm]
    flow[arm] = [len(g),
                 int(g["converted"].notna().sum()),
                 int(g["converted"].notna().sum())]
print(flow.to_string())

for arm, n in PLANNED.items():
    got = int((ab["variant"] == arm).sum())
    flag = "OK" if got == n else "INVESTIGATE"
    print(f"{arm:<10s} planned {n:>5d}  realized {got:>5d}"
          f"  {flag}")

print("\nUnder complete randomization the realized counts "
      "should equal\nthe planned counts exactly. A shortfall "
      "is post-assignment loss:\nreport it, trace it, and do "
      "NOT drop the affected customers.")
```

Expected output: a three-row flow table giving the randomized, outcome-observed, and analyzed counts for each arm; one planned-versus-realized line per arm with an OK or INVESTIGATE flag; and the standing note on what a shortfall means.

Input is the validated file and the registration's declared allocation; the transformation is counting at three stages; the output is page two of the deliverable. The delivery row of Table 11.4 is absent from this cell for an honest reason: the certified ab_test file does not carry a delivery flag, and inventing one would teach students to fabricate the very evidence the flow record exists to demand. Where a messaging log is available, it joins here as an additional reported row — and it stays a reported row. The final print is the cell's real payload, because the instinct a delivery shortfall provokes is exactly the wrong one, and the code says so at the moment the instinct arrives.

Code 11.5. Profile pre-assignment balance

```
cats = ["home_metro", "generation", "loyalty_tier",
        "app_user"]
for col in cats:
    # Represent missingness explicitly. crosstab drops NaN
    # silently, which would hide an imbalance in whether a
    # characteristic was recorded at all.
    profiled = ab[col].astype("string").fillna("Missing")
    share = pd.crosstab(profiled, ab["variant"],
                        normalize="columns")
    share["diff_pp"] = ((share["treatment"]
                         - share["control"]) * 100)
    print(f"\n{col}")
    print(share.round(4).to_string())

nums = ["prior_orders_90d", "prior_spend_90d"]
print("\n", ab.groupby("variant")[nums]
              .agg(["mean", "std", "count"]).round(3))

print("\nRead these as DIAGNOSTICS, not as tests. Chance "
      "imbalance occurs\nunder a valid procedure; a "
      "conspicuous imbalance triggers an audit\nof the "
      "assignment code, not a re-randomization.")
```

Expected output: one composition table per categorical characteristic, each with the treatment and control shares, an explicit Missing row where any values are absent, and the difference in percentage points; then a table of means, standard deviations, and counts for the two numeric characteristics; then the standing note on how to read them.

The categorical block compares each arm's composition share by share; the numeric block compares means and spreads. Every variable used here is measured before assignment, which is what makes the check meaningful — a "balance check" on a post-assignment variable is a treatment effect wearing a diagnostic's costume. Two details are deliberate. Missing values are converted to an explicit "Missing" category rather than being dropped by crosstab, because whether a characteristic was recorded at all can itself differ between arms and is exactly the kind of structured defect a broken randomizer produces. And the closing note is printed inside the cell so that it travels with the output into whatever document the student pastes it into.

VERIFICATION CHECK

Before you run: predict a shape rather than a number. Every categorical share should differ between arms by a fraction of a percentage point, the differences should scatter in both directions with no pattern, and the numeric means should agree closely with standard deviations that match. Predict the realized arm sizes against the registration's declared 5,000 and 5,000, and predict how many "Missing" rows you expect to see.

After you run: check the flow table and the balance profile together, and record the pair as the assignment certificate. It is page two of the deliverable, and Section 11.13 will not let you skip it. Then reason through the three findings that would change what you do next, and note that only one of them stops the lab. Arms of materially unequal size under complete randomization mean customers were lost after assignment, which must be traced and reported — and the affected customers stay in the intention-to-treat comparison while you trace it. A single covariate imbalanced far beyond the others, especially one related to geography or signup date, is the fingerprint of a randomizer keyed to something structured rather than random, and it sends you to the assignment code and its log. And a prior_spend_90d difference of a percent or two between arms is what a chance mechanism produces on ten thousand customers; it is not grounds for discarding the comparison, and if you want the precision back, the instrument is a covariate-adjusted analysis declared in advance, not a re-randomization.

Investigate if: any covariate's difference is large enough that you would have noticed it without looking for it, or if the Missing share differs between arms. The first is an audit of the assignment mechanism. The second is an audit of the data pipeline, and it matters even when the outcome looks fine, because a characteristic recorded differently across arms usually means something else was too.

Code 11.6. Run the sham-split analysis-path check

```
from statsmodels.stats.proportion import proportions_ztest

# NOT an A/A test. This splits COMPLETED control-arm outcomes
# into two sham groups, so it grades the analysis path only:
# the metric, the test statistic, and the code. It touches
# neither the randomizer nor delivery nor tracking.
rng = np.random.default_rng(2026)
ctrl = ab.loc[ab["variant"] == "control",
              "converted"].to_numpy()

for seed in range(5):
    g = np.random.default_rng(2026 + seed)
    sham = g.integers(0, 2, size=ctrl.size)
    counts = np.array([ctrl[sham == 1].sum(),
                       ctrl[sham == 0].sum()])
    sizes = np.array([(sham == 1).sum(), (sham == 0).sum()])
    rates = counts / sizes
    stat, pval = proportions_ztest(counts, sizes)
    print(f"seed {2026 + seed}  sizes {sizes}"
          f"  diff {(rates[0] - rates[1]) * 100:+.2f} pp"
          f"  z {stat:+.2f}  p {pval:.3f}")
```

Expected output: five lines, one per seed, each giving the two sham arm sizes, the difference between their conversion rates in percentage points, the test statistic, and the p-value.

The cell splits the control arm at random into two halves that received identical treatment, then runs the same test the real analysis will run. The true effect here is exactly zero by construction, so the cell grades the analysis path rather than the campaign — and the comment at the top says precisely which parts of the apparatus it does and does not grade, which is the distinction Section 11.5 draws and the reason this guide no longer calls the procedure an A/A test. A genuine A/A test would push both sham arms through the production randomizer, the production send, and the production tracking, and StyleCraft's CRM team owes one before the next experiment rather than after it.

VERIFICATION CHECK

Before you run: predict the output. Each difference should land within a few tenths of a percentage point of zero, and each p-value should be an arbitrary number between 0 and 1 with no particular tendency to be small. Predict how many of the five seeds you expect to return a p-value below 0.05, and write the number down before you look.

After you run: record all five p-values. The lesson is in the expected count rather than in any single line: at a 5 percent threshold, roughly one run in twenty will report a "significant" difference between two groups that received identical treatment, and when it does, nothing in the output will look wrong. That is why a single significant result here — or in a genuine A/A test — is an occasion to investigate rather than a verdict that the infrastructure is broken, and why repeated or systematic failures are the signal that matters. Write that sentence down, because it is the entire intuition behind Section 11.8's peeking and multiple-comparisons failures, and Code 11.10 will turn it into a machine.

Investigate if: the sham arm sizes are wildly unequal, or several of five seeds return small p-values. The first means the sham flag is not doing what you think it is. The second, on a known-zero effect, means the metric or the test call is wrong — which is exactly the defect this cell exists to catch, and catching it here is far cheaper than catching it in the primary analysis.

11.12.3 Lab 11.2, Part A: Sizing the Test That Was Already Run

Power is an owned concept of this chapter and a stated learning objective, and a chapter that asserts a sample-size calculation without running one is teaching a habit it does not model. This part therefore opens where the CRM team opened, before any outcome existed: with the arithmetic that turned a business judgment about the smallest lift worth detecting into a number of customers per arm. Predict the answer first. At a 6 percent baseline, a 1.5-point minimum detectable effect, a 5 percent significance level, and an 80 percent power target, write down how many customers per arm you think the calculation will demand.

Code 11.7. Calculate required sample size and power

```
import numpy as np
from statsmodels.stats.power import NormalIndPower
from statsmodels.stats.proportion import proportion_effectsize

BASELINE = 0.060      # control conversion, from history
MDE      = 0.015      # smallest lift worth detecting: a
                      # BUSINESS judgment, not a statistic
ALPHA    = 0.05
POWER    = 0.80

power_solver = NormalIndPower()

def n_per_arm(baseline, mde, alpha=ALPHA, power=POWER):
    es = proportion_effectsize(baseline + mde, baseline)
    return power_solver.solve_power(
        effect_size=es, alpha=alpha, power=power,
        ratio=1.0, alternative="two-sided")

def power_at(baseline, mde, n, alpha=ALPHA):
    es = proportion_effectsize(baseline + mde, baseline)
    return power_solver.solve_power(
        effect_size=es, alpha=alpha, power=None,
        nobs1=n, ratio=1.0, alternative="two-sided")

need = n_per_arm(BASELINE, MDE)
print(f"required customers per arm: {int(np.ceil(need))}")
print(f"registered per arm:         5000")

print("\nsensitivity: sample size and realized power at 5,000")
print(f"{'MDE (pp)':>9}{'n per arm':>12}{'power @5,000':>14}")
for mde in (0.005, 0.010, 0.015, 0.018):
    print(f"{mde * 100:>9.1f}"
          f"{int(np.ceil(n_per_arm(BASELINE, mde))):>12d}"
          f"{power_at(BASELINE, mde, 5000):>14.3f}")

print("\npower for the planted 1.8-point lift, by arm size")
for n in (1000, 2500, 5000):
    print(f"  n = {n:>5d} per arm -> power "
          f"{power_at(BASELINE, 0.018, n):.3f}")
```

Expected output: the required customers per arm beside the registered 5,000; a four-row sensitivity table giving the required sample size and the realized power at 5,000 per arm for minimum detectable effects of 0.5, 1.0, 1.5, and 1.8 percentage points; and the realized power for the planted 1.8-point lift at 1,000, 2,500, and 5,000 per arm.

Input is four declared numbers, three of them conventions and one of them a business judgment. The transformation is the standard sample-size calculation for a difference in two proportions. The output is the justification for the sample-size row of Table 11.3, plus the sensitivity table that makes the inverse-square relationship visible instead of asserted. Note the structure of the two helper functions: they differ only in which argument is left as None, because a power calculation and a sample-size calculation are the same equation solved for different unknowns, and seeing that in code is worth more than memorizing either formula.

VERIFICATION CHECK

Before you run: predict the required sample size for the registered 1.5-point minimum detectable effect, and predict the shape of the sensitivity table — specifically, predict the ratio between the sample size needed for a 0.5-point effect and the sample size needed for a 1.5-point effect, using the inverse-square rule of thumb from Section 11.5. Then predict the power at 1,000 per arm for the planted 1.8-point lift, and state in one sentence what a value near one-third would mean for a team that ran that test and reported no effect.

After you run: reconcile the figures. The calculation returns 4,379 customers per arm, which the team rounded to 5,000; the 0.5-point row demands 36,778 per arm, roughly eight and a half times the 1.5-point requirement, which is the inverse-square rule showing up slightly softened by the fact that the baseline moves too. At 5,000 per arm the test has 94.5 percent power against the planted 1.8-point lift and 84.9 percent against the 1.5-point effect it was sized for, and only 35.6 percent power at 1,000 per arm. Write the sentence that last number earns: a team running this experiment at a thousand customers per arm would have had a two-in-three chance of missing a real, profitable effect entirely and reporting that the Drop Alert does not work. Then write the second sentence, which is the ethical one: an underpowered experiment still assigns real customers to a control arm, and it buys nothing with their inconvenience.

Investigate if: your required sample size differs from 4,379 by more than a customer or two. The likely cause is a different effect-size convention — some calculators use the raw difference in proportions rather than the arcsine-transformed effect size that proportion_effectsize computes — and the fix is to state which convention you used rather than to assume the other one was wrong.

11.12.4 Lab 11.2, Part B: The Registered Analysis and the Known True Lift

The assignment is certified and the sample is justified. The analysis may now run. It runs exactly once, exactly as registered, at the grain of assignment, on every randomized customer with an observed outcome.

Code 11.8. Estimate absolute and relative lift

```
from statsmodels.stats.proportion import proportions_ztest

# Intention-to-treat: grouped by ASSIGNED variant, with no
# post-assignment filter of any kind applied above this line.
s = (ab.groupby("variant")["converted"]
       .agg(customers="size", conversions="sum"))
s["rate"] = s["conversions"] / s["customers"]
print(s.to_string())

c_n = int(s.loc["control", "customers"])
c_x = int(s.loc["control", "conversions"])
t_n = int(s.loc["treatment", "customers"])
t_x = int(s.loc["treatment", "conversions"])
c_rate, t_rate = c_x / c_n, t_x / t_n

abs_lift = t_rate - c_rate
rel_lift = abs_lift / c_rate
stat, pval = proportions_ztest([t_x, c_x], [t_n, c_n])

print(f"\ncontrol:   {c_x:>5d} / {c_n:>5d} = {c_rate:.4%}")
print(f"treatment: {t_x:>5d} / {t_n:>5d} = {t_rate:.4%}")
print(f"absolute lift: {abs_lift * 100:+.2f} pp")
print(f"relative lift: {rel_lift:+.1%}")
print(f"z = {stat:.2f}   two-sided p = {pval:.4f}")
```

Expected output: the two-row arm summary with customers, conversions, and rate; then the control and treatment rates stated as fractions and percentages, the absolute lift in percentage points, the relative lift as a percentage, and the test statistic with its two-sided p-value.

Input is the certified file at the grain of assignment; the transformation is two proportions and their difference; the output is the estimate and its p-value. Note what the cell does not contain: no subgroups, no secondary metrics, no loop over anything, and no filter on delivery or engagement. That absence is the registration being honored, and it is the most important feature of the code.

Code 11.9. Compute difference and ratio intervals

```
from statsmodels.stats.proportion import (
    confint_proportions_2indep)

# The PRIMARY result: an interval on the DIFFERENCE.
d_lo, d_hi = confint_proportions_2indep(
    t_x, t_n, c_x, c_n, compare="diff", method="wald")
print(f"absolute lift {abs_lift * 100:+.2f} pp"
      f"   95% CI [{d_lo * 100:+.2f}, {d_hi * 100:+.2f}]")

# The relative figure needs its OWN interval, computed as a
# ratio. Rescaling the difference interval by the observed
# control rate treats an estimate as a known constant.
r_lo, r_hi = confint_proportions_2indep(
    t_x, t_n, c_x, c_n, compare="ratio", method="log")
print(f"risk ratio    {t_rate / c_rate:.3f}"
      f"        95% CI [{r_lo:.3f}, {r_hi:.3f}]")
print(f"relative lift {rel_lift:+.1%}"
      f"      95% CI [{r_lo - 1:+.1%}, {r_hi - 1:+.1%}]")

# The shortcut, printed only so its error is visible.
print(f"\nWRONG (difference CI rescaled by control rate): "
      f"[{d_lo / c_rate:+.1%}, {d_hi / c_rate:+.1%}]")

# Both standard errors, for the record.
se_un = np.sqrt(c_rate * (1 - c_rate) / c_n
                + t_rate * (1 - t_rate) / t_n)
pool = (c_x + t_x) / (c_n + t_n)
se_pool = np.sqrt(pool * (1 - pool) * (1 / c_n + 1 / t_n))
print(f"SE unpooled {se_un * 100:.4f} pp"
      f"   SE pooled {se_pool * 100:.4f} pp")
```

Expected output: the absolute lift with its 95 percent interval on the difference; the risk ratio with its 95 percent interval; the relative lift with the interval implied by the ratio interval; the interval the rescaling shortcut would have produced, labeled as wrong; and the unpooled and pooled standard errors.

Input is the four counts from Code 11.8; the transformation is two different interval constructions on the same data; the output is the pair of intervals the deliverable reports and the one it must not. The ratio interval is computed on the log scale and exponentiated, which is the standard construction and the reason the interval is asymmetric around the point estimate. Printing the shortcut beside the honest answer is deliberate pedagogy rather than clutter: students who have never seen the two side by side tend to assume the difference is enormous, and the more useful lesson is that it is modest here and unpredictable in general — which is exactly why the shortcut cannot be trusted rather than merely tolerated.

VERIFICATION CHECK

Before you run: predict the numbers in full, because this is the chapter's verification theme. The control rate should land near 6.0 percent, since that is the baseline the power calculation was built on; the treatment rate near 7.8 percent; the absolute lift near 1.8 percentage points, because the dataset's published design note states the planted true lift and the whole purpose of a correct analysis is to recover it. Predict the interval's width too, from the arm sizes: with 5,000 per arm at rates near 6 and 8 percent, the standard error of the difference is close to 0.51 percentage points, so the 95 percent interval is roughly two points wide. Predict where the relative-lift interval will fall, and predict whether the rescaling shortcut will look too wide or too narrow.

After you run: grade four things. Does the point estimate land near 1.8? Does the interval contain 1.8? Is the interval about as wide as you predicted — an interval far narrower than two points is the grain slip of Section 11.11, so confirm the analysis counted customers rather than orders, and an interval far wider means the arms are smaller than the registration promised, which sends you back to Code 11.4. And is the relative interval, running from roughly 12 to 50 percent, visibly wider than the 13-to-47 the shortcut produced? Then write the two sentences the deliverable opens with: the control and treatment rates stated plainly, and the lift stated in percentage points with its interval attached and the relative figure second. Finally, compare the interval against the observational 5.30 points from Part A and state the comparison precisely, because the draft of this chapter stated it wrongly: the observational difference is nearly three times the experimental point estimate of 1.80 points and nearly twice the upper end of the experimental interval. Those are two different multiples of two different quantities, and quoting the larger one against the smaller reference is the kind of small inaccuracy that a finance partner will find.

Investigate if: the point estimate is far from 1.8, or the interval excludes it. Neither should happen on the certified file, and both are findings about the pipeline rather than about StyleCraft. Check first that the grouping variable is the assigned variant and not something derived from behavior after assignment, then that no filter was applied between Code 11.3 and Code 11.8.

11.12.5 Lab 11.2, Part C: Manufacturing a False Winner

Section 11.8 asserted that peeking inflates false positives far beyond the nominal 5 percent, and gave two numbers rather than one. This part makes the student produce both, on StyleCraft's real data, in a setting where the true effect is known to be exactly zero because both sham groups come from the control arm.

Code 11.10. Simulate repeated peeking

```
import numpy as np
from statsmodels.stats.proportion import proportions_ztest

ctrl = ab.loc[ab["variant"] == "control",
              "converted"].to_numpy()
LOOKS, ALPHA, TRIALS, MIN_PER_ARM = 14, 0.05, 1000, 100

def one_trial(y, g, keep_path=False):
    """One sham experiment examined LOOKS times. Returns the
    verdict under three stopping policies -- final (look once
    at close), any (stop at any significant difference), and
    favor (stop only when significant AND treatment ahead) --
    plus, optionally, the p-value at every look."""
    order = g.permutation(y.size)
    y, arm = y[order], g.integers(0, 2, size=y.size)
    cuts = np.linspace(y.size / LOOKS, y.size,
                       LOOKS).astype(int)
    final = any_hit = favor_hit = False
    path = []
    for k in cuts:
        a, b = y[:k][arm[:k] == 1], y[:k][arm[:k] == 0]
        if min(a.size, b.size) < MIN_PER_ARM:
            path.append(np.nan)
            continue
        _, p = proportions_ztest([a.sum(), b.sum()],
                                 [a.size, b.size])
        path.append(p)
        final = p < ALPHA
        if p < ALPHA:
            any_hit = True
            favor_hit = favor_hit or a.mean() > b.mean()
    return final, any_hit, favor_hit, path if keep_path else None

g = np.random.default_rng(11)
res = np.array([one_trial(ctrl, g)[:3] for _ in range(TRIALS)])
print(f"{TRIALS} sham experiments, {LOOKS} looks each, "
      f"true effect exactly 0")
for i, label in enumerate(("FINAL look only          ",
                           "ANY look, either direction",
                           "ANY look, treatment ahead ")):
    print(f"  significant at {label}: {res[:, i].mean():.1%}")
print(f"  1 - 0.95**{LOOKS} (independent-look bound): "
      f"{1 - 0.95 ** LOOKS:.1%}")
```

Expected output: a header line stating the number of sham experiments, the number of looks, and that the true effect is exactly zero, then the three false-positive rates under the three stopping policies and the independent-look upper bound for comparison.

The function shuffles the control arm into a random arrival order, splits it at random into two sham arms, and walks forward through the data computing the same two-proportion test at fourteen evenly spaced moments — the analyst checking her dashboard each morning. Three counters record three different analysts. The disciplined one looks once at the end. The second stops at any significant difference in either direction. The third, who is the realistic one, stops only when the result is significant *and* the treatment arm is ahead, which is what actually

happens in organizations because nobody halts a test to announce that the new idea is losing. Separating the second from the third is the correction this chapter's draft required: the simulation used to count either-direction hits while the prose described a treatment-is-winning policy, and those are different numbers answering different questions. The optional path return, which Code 11.11 switches on, then makes the phenomenon watchable rather than merely tabulated.

Code 11.11. Trace one analyst's p-value path

```
# The aggregate rates say how often peeking misleads. This
# says what it LOOKS like from the inside: one sham
# experiment, examined fourteen times, printed look by look.
g_one = np.random.default_rng(7)
final, any_hit, favor_hit, path = one_trial(
    ctrl, g_one, keep_path=True)

print("look   n per arm (approx)   p-value   would stop?")
for i, p in enumerate(path, start=1):
    approx = int(ctrl.size * i / LOOKS / 2)
    if np.isnan(p):
        print(f"{i:>4}   {approx:>16}   {'--':>7}   too small")
    else:
        stop = "YES <--" if p < ALPHA else ""
        print(f"{i:>4}   {approx:>16}   {p:>7.3f}   {stop}")

print(f"\ndisciplined analyst (final look only): "
      f"{'declares a difference' if final else 'declares nothing'}")
print(f"undisciplined analyst (any look)      : "
      f"{'declares a difference' if any_hit else 'declares nothing'}")
```

Expected output: a fourteen-row table giving the look number, the approximate customers per sham arm at that look, the p-value, and a flag marking any look at which an undisciplined analyst would have stopped; then one line each for what the disciplined and the undisciplined analyst would have concluded from the same experiment.

Input is the same function, called once with its optional path return switched on. The transformation is nothing; the output is the exhibit. Read the p-value column downward and watch it wander: on some seeds it dips below 0.05 in the middle of the window and climbs back out, which is the entire phenomenon in one column of numbers. The last two lines are the cell's argument. Two analysts, one experiment, identical data, opposite conclusions — and the only difference between them is a rule written down before launch. Run the cell across several seeds and note how often the two lines disagree.

VERIFICATION CHECK

Before you run: predict all three rates. The final-look rate should land near 5 percent, because that is what the threshold means and the data contain no effect — this number validates the simulation itself, so if it is far from 5 the rest is not interpretable. Then predict the any-look rate. The arithmetic of fourteen independent tests would suggest 1 − 0.95 to the fourteenth, about 51 percent, but the looks are not independent because each contains the previous one, so the true inflation is smaller. Predict where it lands, and then predict the favorable-only rate relative to it — the answer is close to half, and it is worth reasoning out why before you see it.

After you run: record all three. On this design the final-look rate sits near 5 percent, the any-direction rate near 22 percent, and the treatment-ahead rate near 11 percent; across seeds the three land in roughly the 4-to-6, 20-to-24, and 10-to-12 percent bands, so quote a band rather than a decimal. Then read the p-value path printed by Code 11.11 and describe what an analyst would have seen: a number wandering, dipping below 0.05 somewhere in the middle on some seeds, and climbing back out. That wandering is the entire phenomenon. Now write the three sentences the exhibit has earned. First, in a test with no effect whatsoever, roughly one analyst in nine who checks daily and stops when her treatment is ahead will ship a winner that does not exist. Second, that winner's report will look completely normal — a p-value below 0.05, an interval excluding zero, a plausible lift — because nothing in the final output records how many times the data were examined. And third, the defense is not statistical sophistication but the stopping rule in Table 11.3, declared before launch and honored afterward, which costs nothing and is skipped constantly.

Investigate if: the final-look rate is far from 5 percent, or the treatment-ahead rate is not roughly half the any-direction rate. The first means the simulation or the test call is wrong and no other number in the cell can be trusted. The second would be genuinely strange under a true null, because a symmetric null should produce significant results in both directions about equally often — so check that the arm comparison and the mean comparison use the same arm as arm one.

11.12.6 Lab 11.2, Part D: Practical Significance, Temporal Stability, and the Rollout Decision

The estimate is credible and both intervals are honest. This part converts them into the decision, stress-tests the economics that carry the conversion, and then applies the one durability diagnostic the certified file can actually support.

Code 11.12. Price the interval against break-even

```
# The DOSE comes first: the break-even line is undefined
# without it. Table 11.3 declares one alert per treated
# customer, so cost per assigned customer is one exposure.
COST_PER_ALERT   = 0.12
ALERTS_PER_CUST  = 1
CONTRIB_PER_CONV = 28.00

cost_per_cust = COST_PER_ALERT * ALERTS_PER_CUST
break_even = cost_per_cust / CONTRIB_PER_CONV
print(f"dose: {ALERTS_PER_CUST} alert(s) -> cost per assigned "
      f"customer ${cost_per_cust:.2f}")
print(f"break-even lift: {break_even * 100:.2f} pp")

print("\nvalue per 1,000 customers enrolled, at this dose")
for label, lift in (("interval low", d_lo),
                    ("point estimate", abs_lift),
                    ("interval high", d_hi)):
    incr = lift * 1000
    net = incr * CONTRIB_PER_CONV - 1000 * cost_per_cust
    print(f"  {label:<14s} lift {lift * 100:5.2f} pp"
          f"  {incr:5.1f} conversions  net ${net:8.2f}")
print(f"\ndecision rule met (lower bound > break-even): "
      f"{d_lo > break_even}")

# The line is an ASSUMPTION, so vary it. Contribution is an
# average over observed orders; incremental converters need
# not match it, and the rollout dose may differ.
print("\nbreak-even lift (pp) across dose and contribution")
print(f"{'contribution':>13}", end="")
for lab in ("1 alert", "2 alerts"):
    print(lab.rjust(9), end="")
print()
for contrib in (20.00, 24.00, 28.00, 32.00):
    print(f"{'$' + format(contrib, '.2f'):>13}", end="")
    for k in (1, 2):
        be = COST_PER_ALERT * k / contrib
        mark = "" if d_lo > be else "  <-- inside CI"
        print(f"{be * 100:>9.2f}{mark}", end="")
    print()
```

Expected output: the declared dose with the implied cost per assigned customer and the break-even lift; the incremental conversions and net contribution per thousand customers at the interval's lower bound, the point estimate, and the upper bound; whether the registered decision rule is met; and a sensitivity grid of break-even lifts across four contribution figures and two doses, flagging any cell the confidence interval does not clear.

Input is the interval from Code 11.9 plus three business constants; the transformation is the arithmetic of Section 11.7; the output is the decision rule declared in Table 11.3, evaluated, plus the sensitivity that keeps the evaluation honest. Note that the constants are inputs to the analysis and belong in the registration, not numbers chosen after the lift was known — an analyst who selects the contribution figure after seeing the estimate has invented a new way to peek. The sensitivity grid is the cell's real contribution to the deliverable, because it answers the question a

finance partner asks second: not "does it pay?" but "how wrong would our cost and margin assumptions have to be before it stops paying?" On these numbers the answer is reassuring at one alert and much less so at two, and the flagged cells are where the conversation goes.

Code 11.13. Assess temporal stability

```
# NOT a novelty test. Customers assigned in the first and
# second halves of the window are different cohorts, exposed
# to different drops under different calendar conditions, so a
# gap between halves is not evidence of decay with exposure.
# Novelty needs repeated exposures for the SAME assignment,
# which this one-outcome-per-customer file does not contain.
span = ab["assign_date"].max() - ab["assign_date"].min()
half = ab["assign_date"].min() + span / 2
ab["cohort"] = np.where(ab["assign_date"] <= half,
                        "first half", "second half")

for cohort, grp in ab.groupby("cohort"):
    gs = (grp.groupby("variant")["converted"]
             .agg(n="size", x="sum"))
    r = gs["x"] / gs["n"]
    lift = (r["treatment"] - r["control"]) * 100
    per_arm = int(gs["n"].min())
    se = np.sqrt(sum(p * (1 - p) / n for p, n
                     in zip(r, gs["n"]))) * 100
    print(f"{cohort:<12s} n/arm {per_arm:>5d}"
          f"  control {r['control']:.2%}"
          f"  treatment {r['treatment']:.2%}"
          f"  lift {lift:+.2f} pp  (SE {se:.2f})")

print("\nRead the two SEs before reading the two lifts: at "
      "roughly half\nthe arm size, each half-window estimate "
      "is far too imprecise to\ndetect a 1.8-point effect, so "
      "the halves WILL differ by luck.\nThis diagnostic can "
      "flag a difference big enough to investigate.\nIt cannot "
      "certify stability, and it cannot measure novelty.")
```

Expected output: one line per assignment half giving the per-arm sample size, the control and treatment rates, the estimated lift in percentage points and its standard error; then the standing note on what the diagnostic can and cannot establish.

The cell splits the experiment by when customers were assigned and re-estimates the lift in each half. Three things about it are corrections to the draft of this chapter. It is named a temporal-stability diagnostic rather than a novelty check, because comparing cohorts assigned on different dates cannot isolate a response to unfamiliarity. It prints each half's standard error beside its lift, so the reader sees the imprecision before the difference rather than after. And its closing note states plainly that the file's single binary outcome per customer contains no repeated-exposure information at all, which is why the rollout plan must commission the multi-exposure design Section 11.8 describes rather than treating this exhibit as a substitute for it.

VERIFICATION CHECK

Before you run: predict the two half-window lifts and, more importantly, predict their standard errors. With roughly 2,500 customers per arm per half, the standard error of each half's difference is close to 0.72 percentage points, which means the two halves can differ by a point and a half without anything having happened. Write down, before you look, what difference between halves would actually warrant investigation given that imprecision.

After you run: compare the two lifts against the two standard errors, and write the interpretation in one sentence: the half-window estimates are too imprecise to detect novelty decay, so this diagnostic can only flag a difference dramatic enough to demand investigation and cannot certify its absence. Then state what the rollout plan must therefore carry — a scheduled re-measurement, and a design that follows one randomized assignment across several consecutive drops so that a treatment-by-exposure pattern is estimable at all. Finally, note what this whole part has done to the CMO's proposal: it supports expanding the Drop Alert at the tested dose, and says nothing whatever about defunding paid social, per Section 11.7's third restriction.

Investigate if: the two halves differ by more than about three standard errors, or either half's control rate is far from 6 percent. The first is large enough to be worth tracing to a calendar event, a promotion, or a delivery incident in one half of the window. The second suggests the assignment window overlapped something — a holiday, a site outage, a competing campaign — that the registration did not anticipate, which belongs in the deliverable's deviations list.

11.12.7 Lab 11.2, Part E: The Round Trip and the Rollout Paragraph

The last part is the assistant round trip of Section 11.11, run end to end, and it is where the chapter's deliverable is actually written. Work in four moves, documenting each per Appendix D.

First, the registration prompt. Draft the preregistration for a *second* experiment — the one Section 11.7's third restriction says StyleCraft needs, a paid social incrementality test with a geographic holdout — using the template of Table 11.3 and the instruction that the assistant must write UNSPECIFIED and ask rather than filling gaps. Count the UNSPECIFIED rows it returns, and check specifically whether it left the dose and the guardrail-computability rows unspecified or invented them. That count is a measurement of how much of your own design was missing from your prompt, and it is the most useful number the assistant will produce all week.

Second, the corrupted-analysis audit. Hand the assistant the ab_test file with a deliberately misleading prompt: ask it to "check whether the campaign worked, and see which customer segments responded best." Then audit what comes back against Table 11.7. Expect four specific failures, and record whether each occurred: a segment harvest with unadjusted p-values across home_metro, generation, and loyalty_tier; a headline built on whichever subgroup won; no mention anywhere that the subgroups were not registered; and no statement of the treatment dose the reported lift belongs to. The exercise is not a trick. It is a demonstration that the assistant answers the question it was asked, and that the analyst's job is to ask the registered question.

Third, the adversarial prompt. Ask the assistant to argue against your own result: every rival explanation, every contamination path, every difference between the tested population and the

rollout population, and every way a rollout at two alerts per week differs from a test at one. Keep the items that survive your judgment and place them in the deliverable's limitations section, cited to nothing — they are your limitations now.

Fourth, the drafting prompt, with only the audited numbers pasted in.

VERIFICATION CHECK

Before you run: write down what you expect the assistant to get wrong, item by item, against Table 11.7's three supplemental points. Predicting the failure modes before you see them is what turns the round trip from a demonstration into an audit, and it is the part students most often skip.

After you run: run every drafted paragraph against three tests. The verb test, per Section 11.10: the paragraph may say the Drop Alert *caused* a lift among the tested population at the tested dose within the stated interval; it may not say the rollout *will* lift conversion by 30 percent. The number test: every figure in the paragraph must appear in a lab output above, with the absolute lift stated before the relative one, the control rate beside it, and the relative interval computed as a ratio rather than rescaled. The qualification test: the paragraph must name the tested population, the tested window, and the tested dose, because dropping any of the three converts a defensible finding into the deck that opened this chapter. Then write the final version yourself. The deliverable's key sentence, in the pattern Section 11.13 will polish: among SMS-opted-in customers, over a fourteen-day window, at one Drop Alert per customer, the alert raised conversion from 6.0 percent to 7.8 percent — a lift of 1.8 percentage points, 95 percent interval 0.8 to 2.8 points, or roughly $106 to $662 of contribution per thousand customers per drop window against a break-even of 0.43 points — and the program therefore pays for itself across the full range of effects the experiment supports, subject to the stated cost and contribution assumptions, the extrapolation limits, and the re-measurement schedule below.

Investigate if: the assistant's draft is clean on the first pass. That is possible and it is not a reason to relax, because the failure modes of Section 11.11 are silent rather than loud. Check the two that hide best: whether any reported interval is a rescaled difference interval wearing a ratio's label, and whether the sample size implied by the interval's width matches the number of randomized customers rather than the number of orders.

11.13 Marketing Interpretation and Managerial Insight

The labs produced a certificate, two intervals, and a break-even comparison; the reallocation meeting runs on sentences, and the sentences an experiment produces are unusually easy to say wrong. This section translates, and it does so partly by exhibiting the wrong readings and correcting them, because experimental results attract a particular family of misinterpretation: they arrive with the authority of a trial, which makes stakeholders more rather than less willing to stretch them past what was tested.

The first wrong reading is the one this chapter's outline predicted, and it will be spoken within a minute of the result going up: "It's significant at p less than point-oh-five, so the campaign will lift sales 12 percent at scale." Five errors are packed into a single clause, and each has a section behind it. The p-value has been read as a statement about the future and about

magnitude, when it is neither — it is a measure of how surprising this difference would be if the true effect were zero, and Section 7.10's warnings apply unchanged (Wasserstein & Lazar, 2016). The 12 percent has arrived from nowhere; the experiment measured a 30 percent relative lift on conversion, with an interval running from roughly 12 to 50 percent, and conversion is not sales — revenue per assigned customer was a declared *secondary* metric precisely because a treatment can raise conversion while lowering average order value, and the deliverable reports it as exploratory rather than promoting it to the headline. The words "at scale" quietly change the population from SMS-opted-in customers with thirty-day accounts to whatever the rollout reaches, which the experiment did not test. The claim carries no dose, so it silently transports a one-alert result onto a rollout the CRM team wants to run at two, which is a different intervention with twice the cost. And "will" has replaced "caused," which converts a bounded past-tense measurement into a promise, exactly the move Section 10.15 called laundering when a forecast's interval was stripped in the retelling. The corrected sentence is longer and worth the length: among SMS-opted-in customers, over a fourteen-day window, at one Drop Alert per customer, the alert caused conversion to rise 1.8 percentage points from a 6.0 percent base — with the evidence consistent with anything from 0.8 to 2.8 points — and a chain-wide quarter-long rollout at higher frequency would be expected to deliver somewhat less per alert, for reasons the limitations section names.

The second wrong reading is quieter and destroys more experiments than the first, because it ends programs instead of overselling them: "The test wasn't significant, so the campaign doesn't work." An experiment that fails to reject is not an experiment that proves the null, and the distinction is entirely a matter of power. Code 11.7 makes the point in numbers rather than in principle: a test with 1,000 customers per arm, run on this same genuinely valuable 1.8-point treatment, would have found it only about a third of the time, and a test at 500 per arm could only have reliably detected a lift of nearly 5 points. Reporting either outcome as "no effect" retires a profitable program on the strength of an experiment that was never capable of finding it. The standing rule this guide adopts: a null result is reported with its minimum detectable effect attached. "The test found no significant difference; it was powered to detect a lift of 3 points or more, so effects smaller than that remain entirely possible" is an honest sentence, and it is a different recommendation — usually "test again, larger" rather than "stop."

The third wrong reading belongs to the subgroups, and it is the most seductive because it feels like insight: "The overall lift was modest, but look — it was 6.2 points among Gen Z suburban app users. Let's target them." Section 11.8 named the mechanism, and the deliverable's defense is structural rather than rhetorical. Subgroup results appear in an exploratory section, visually separated, with a standing note that they were not registered, that many were examined, and that the largest of many estimates is biased upward by the act of being the largest. What they license is a next experiment, registered in advance on that subgroup, which is a cheap and excellent use of the finding. What they do not license is a targeting decision, and an analyst who lets an unregistered subgroup drive a rollout has spent the credibility the whole design was built to earn.

The deliverable that survives all three misreadings has a fixed anatomy, assembled entirely from lab outputs, and its ordering is itself an argument. Page one is the registration, dated, with the treatment dose in the same table as the primary metric and any deviations from the plan listed as deviations — not because deviations are disqualifying but because undisclosed ones are. Page two is the assignment certificate from Lab 11.1 Part B: the assignment-flow record from randomization through outcome observation, realized arm sizes against the declared allocation, the covariate balance profile read as a diagnostic, the sham-split analysis-path check with a note that a genuine end-to-end A/A test remains outstanding, and the contamination paths named with their expected direction. Only then, page three, the primary result: control rate, treatment rate, absolute lift with its interval on the difference, relative lift second with its own interval on the ratio, and the same difference interval restated in incremental conversions and dollars per thousand customers at the tested dose. Page four is the decision: the break-even lift computed from the program's own dose and economics, the comparison of the full interval against that line, the sensitivity of that line to contribution and dose, and the recommendation stated as a rule rather than a hope — expand at the tested dose, because the program pays across the entire interval under the stated assumptions, with the asymmetric-cost reasoning of Section 2.7 made explicit. Page five is the limitations, which for an experiment are unusually specific and unusually important: the tested population against the rollout population, the fourteen-day window against a quarter, the one-alert dose against the proposed frequency, the novelty risk with the multi-exposure design that would actually measure it, and the sentence the CMO will not enjoy — that the reallocation from paid social requires an incrementality estimate for paid social that does not exist, and that the correct next step is the experiment which would produce it. Page six is the guardrails and the reversal plan: the revenue-per-assigned-customer threshold declared in Table 11.3, who monitors it, what happens if it trips, and the instrumentation obligation that would let the next test declare an unsubscribe guardrail it can actually compute. And the exploratory section, last and labeled, holding the subgroups and secondary metrics as candidates for the next registration.

One further translation deserves its own paragraph, because it is what makes this chapter's deliverable different in kind from the four before it. Chapters 8, 9, and 10 produced artifacts whose value was the number inside them. This chapter's artifact is valuable mostly for a decision it makes possible in the *other* direction: the cheapest thing an experiment does is kill a bad idea for the price of a test rather than the price of a quarter. StyleCraft's Drop Alert survived. Had the experiment returned an interval from −0.4 to +0.9 points, straddling the break-even line, the deliverable would have saved a four-hundred-thousand-dollar reallocation for the cost of ten thousand messages, and that is the outcome an experimentation practice earns its budget on. Managers judge experiments by the winners they find. Analysts should judge them by the losers they catch, and should say so in the meeting, because it is the argument that funds the second experiment.

11.14 Business Analytics in Practice

This section turns from the fictional case to how controlled experimentation operates — and fails — in professional organizations. It looks at where experiments have become the default evidence standard at the largest technology companies, where their absence has quietly distorted a decade of advertising budgets, and where the discipline that looks pedantic in a textbook turns out to be the entire difference between a testing program and a rumor mill. The first two vignettes draw on published research and documented industry practice; the third is a composite of recurring commercial testing patterns rather than a report about a single named organization.

11.14.1 Experimentation Culture as Infrastructure

The first vignette is the experimentation culture, which at the companies that pioneered it is now a piece of infrastructure rather than a practice. Microsoft's Bing organization, Booking.com, Amazon, and their peers run controlled experiments continuously and at a scale that is difficult to picture from a marketing department: thousands of concurrent tests, a substantial fraction of all product and messaging changes shipped only after passing one, and platforms that handle assignment, delivery, and analysis so that the analyst's judgment is spent on design rather than plumbing (Kohavi et al., 2013; Kohavi et al., 2020).

Three of the practices these organizations converged on are directly borrowable by a team of four with no platform at all. The first is a single declared overall evaluation criterion — one metric that decides, chosen in advance, resisting the pull toward a dashboard of twelve — which is the preregistration's primary-metric row operating as culture rather than paperwork. The second is guardrail metrics: measures that are not expected to improve but must not degrade, such as unsubscribe rate, page latency, or margin, checked on every test so that a conversion win bought at the cost of a churn increase is caught at the test rather than at the quarterly review. The corollary that matters for a small team is instrumentation: a guardrail is only a guardrail if the pipeline delivers the field, which is why mature programs treat "we could not compute that" as a defect with an owner rather than a footnote. The third, and the most transferable, is the standing A/A test: the same organizations that run thousands of experiments run continuous null experiments through the full production pipeline, because a randomizer that silently breaks invalidates everything downstream of it and the only way to know is to keep asking a question whose right answer is already known. Note that this is the end-to-end version, not the analysis-path check a spreadsheet can do — the point is that the assignment, the send, and the tracking are all in the loop. The practice lesson for a small team is proportion, not scale. Nobody needs a platform to declare one metric, name a guardrail the pipeline can actually measure, and split a list in half through the real send machinery twice a year to check that the splitter still works.

11.14.2 Advertising Incrementality and the Measured-Channel Bias

The second vignette is advertising incrementality, and it is the one with the largest number attached. The industry standard for evaluating digital advertising has long been attribution: assigning credit for conversions to the ads that preceded them, using data the platforms

themselves supply. Attribution is observational, and Section 11.3 explains exactly what is wrong with it — ads are shown to people the targeting system predicts are most likely to convert, so many of the conversions that follow would have happened anyway, and the resulting measurement credits the advertiser for its own customers' intentions. This is the upward-pointing case of selection bias, and it is upward for a specific and identifiable reason: the selection mechanism targets high-propensity customers.

When experiments have been run against that standard, the gap has been large. eBay's large-scale field experiments on branded paid search found that returns were far lower than the attributed figures suggested, with much of the measured value reflecting traffic that would have arrived through other channels (Blake et al., 2015). Lewis and Rao (2015) established the structural reason the industry rarely notices: sales outcomes are so variable relative to advertising's true effect that credibly measuring the return requires experiments far larger than firms typically run, which means most advertising ROI numbers in circulation come from tests without the power to produce them. Gordon et al. (2019) closed the loop by running experiments and observational analyses on the same campaigns with rich covariates available, and found the observational estimates diverged from the experimental ones by margins large enough to reverse budget decisions — and, importantly for Section 11.3's argument about direction, diverged in both directions rather than consistently upward. The industry's response is the geo experiment and the holdout: randomizing markets rather than individuals when the medium cannot be targeted, or withholding advertising from a randomly chosen share of the audience and comparing (Vaver & Koehler, 2011).

The practice lesson lands on Section 11.7's third restriction, and it is the most valuable sentence a junior analyst can say in a budget meeting. A marketing organization that has measured one channel experimentally and the rest by attribution will systematically move money toward the measured channel — not because it performs better but because it is the only one whose number was ever tested. The bias is not in any single estimate. It is in the comparison, and it favors whichever program was brave enough to be measured.

11.14.3 The Testing Program That Manufactured Knowledge

The third vignette is smaller, is included because most students will encounter its version before they encounter either of the others, and is presented as a composite vignette built from recurring commercial testing practices rather than as a report about a specific firm. A mid-size retailer's CRM team runs subject-line tests on its weekly newsletter, checking results on a dashboard each morning and declaring a winner as soon as one appears, which typically happens within a day or two. Over a year the program accumulates an impressive log of victories — dozens of winning lines, each with a lift of eight to fifteen percent — and the aggregate newsletter conversion rate does not move at all. When the team eventually re-runs a sample of its past winners as fresh, properly sized, single-look tests, most of them reproduce nothing.

The mechanism is Section 11.8's first failure, operating exactly as Code 11.10 stages it: with a small daily audience and a stopping rule of "stop when it wins," the program is not testing

subject lines but sampling noise until the noise is flattering, and the asymmetry of only ever stopping on wins converts an unbiased random process into a machine that produces a winner far more often than the nominal threshold suggests (Johari et al., 2017). The costly part is not the wasted year. It is that the team has built a set of beliefs about its own customers — what tone works, what urgency works, what emoji works — out of pure noise, and those beliefs shape creative decisions well beyond the newsletter. The practice lesson: a testing program without a stopping rule does not merely fail to learn, it actively manufactures false knowledge, and it does so faster the more tests it runs.

11.14.4 In Your First Analyst Job

In your first analyst job, these vignettes compress into one expectation, and it is this chapter's closing thread: the cheapest experiment stops a bad idea before rollout. The glamorous version of experimentation is the winner — the change that lifted conversion, the test that justified the budget, the slide with the green arrow. The version that pays for the practice is the other one: the redesign that tested flat and did not ship, the offer that tested negative and was withdrawn, the reallocation that a ten-thousand-customer test showed was resting on a number three times too large. Those results generate no celebration and no slide, and they are where nearly all of the return lives, because the cost of testing an idea is always a small fraction of the cost of rolling one out. The analyst who understands this argues for experiments differently from the one who does not. She does not promise that the test will find a winner, which she cannot know. She promises that the organization will find out cheaply either way — and that promise, unlike the other one, is one she can keep every time.

11.15 Ethics, Testing on Customers, and the Fairness of the Holdout

The Business Analytics in Practice section ended on experiments as a way to fail cheaply; this section examines what it means that the failing is done to people who did not volunteer. It extends the guide's running discussions — data use (Section 1.13), problem framing (Section 2.12), measurement design (Section 3.13), cleaning as editorial power (Section 4.14), honest summarization (Section 5.14), differential treatment of segments (Section 6.16), causal language (Section 7.15), acting on predictions about people (Section 8.15), fairness across groups (Section 9.15), and forecast accountability (Section 10.15) — to what experimentation adds, which is a category the previous chapters did not contain. Every prior chapter analyzed customers. This chapter's methods do something to them, deliberately, by a chance mechanism, and specifically in order to see what happens.

Start with the fact that makes the ethics unavoidable rather than optional, and state it precisely, because the loose version of the sentence gets the ethics backward. An experiment knowingly assigns customers to different experiences while the organization remains genuinely uncertain which is better overall. It does not assign anyone to a worse experience on purpose — if the company already knew which arm was better, there would be nothing to learn and no ethical basis for withholding it. Some customers will retrospectively turn out to have received the

inferior option, and that is a real cost, but it is not knowable at the moment of assignment, and the uncertainty is precisely what makes the test defensible rather than what makes it suspect. The control arm in StyleCraft's test receives the existing standard drop announcement — the experience every customer received last month and most customers still receive — not a degraded one. And the customer in the control arm was not asked. She agreed, somewhere in a terms-of-service document she did not read, to receive marketing communications; she did not agree, in any sense the word normally carries, to be a comparison group.

The reference case for what happens when this is taken lightly is the emotional contagion study, in which the news feeds of hundreds of thousands of users were altered to show more or less emotionally positive content and the effect on their own posting was measured (Kramer et al., 2014). The research was competent, the effect was real, the manipulation was small, and the publication triggered a public and academic reaction severe enough that the journal issued an editorial expression of concern about the consent procedures (Verma, 2014). The instructive part for a marketing analyst is not the scientific ethics debate but the commercial one. What the public objected to was not the size of the effect; it was the discovery that the emotional content of their feed was something a company would deliberately vary to see what happened to them. The line the episode drew is not about statistical harm. It is about the difference between a company improving its product and a company running an experiment on its users' emotional states, and that difference is legible to ordinary people even when it is invisible in the analysis plan.

This suggests the working distinction the guide adopts, which is coarse but usable. Testing which of two versions of a product experience serves customers better is ordinary business improvement, the kind every company has always done informally and which a controlled experiment merely does honestly. Testing something whose mechanism is the customer's psychology rather than the product's quality — manipulating scarcity signals, urgency, emotional framing, or the perception of price — is a different activity wearing the same clothes, and it deserves a harder question before it runs: would we be comfortable explaining this test, in plain language, to the customers in it? The question is not a bright line and it is not meant to be. It is a prompt that reliably separates the tests nobody will mind from the tests that end up in a journalist's inbox, and an analyst who asks it out loud in the design meeting has done the most useful thing available.

Marketing's own hardest case is not emotional framing but price and offer testing, and it deserves naming because students will meet it. Randomizing which customers see which discount is a real experiment with real value, and it also means charging different people different amounts for the same item on the basis of a chance draw. The practices that keep this defensible are specific. Test discount *depth* on offers rather than base prices, so nobody pays more than the posted price. Keep the treatment duration short and the arms' economic difference modest. Honor the losing arm afterward where the difference was material, which is the practice of extending the winning offer to the control group once the test closes; that practice is not required by any statute and is the clearest available signal that an organization understands what

it just did. And handle protected characteristics with the precision the fairness discipline of Section 9.15 established, which the draft of this chapter did not. The rule is not that the assignment variable must never correlate with a protected characteristic — a valid chance mechanism will show small realized correlations with everything, by luck, and demanding zero correlation would forbid randomization itself. The rule is that the assignment mechanism must not use protected or proxy characteristics as inputs, and that the realized distribution of benefits, burdens, offers, and error rates should be audited across relevant groups after the fact, with a stated response if the audit finds a pattern.

Which brings the section to the holdout, and to the question the chapter poses most sharply: who is denied the benefit, and why is that defensible?

CONCEPT

The Holdout Is a Debt You Repay

A control arm is customers who did not receive something the company hopes is good, held there so the company can find out whether it actually is. The arrangement is defensible under four conditions, and the analyst should be able to state all four before the test launches. First, genuine uncertainty: the treatment's value is unknown, which is precisely why the test is being run — withholding a benefit already known to be beneficial is not an experiment but a cost saving with a scientific alibi. Second, proportion: the holdout is no larger and no longer than the power calculation requires, which is the ethical function of Code 11.7's arithmetic — an oversized or open-ended holdout denies more people more benefit than the question needs, and an undersized one denies them the benefit and learns nothing, which is worse.

Third, bounded harm: the control arm receives the existing experience rather than a degraded one, and any guardrail that would signal real damage is declared in advance, computable from the data the pipeline actually delivers, and attached to a rule for stopping. Fourth, repayment: when the treatment wins, the holdout receives it, promptly, and the plan to do so exists before the test starts rather than being remembered afterward. Under those four conditions the holdout is the cost of finding out, borne briefly and repaid — and an organization that cannot meet all four should notice that its discomfort is information about the test rather than squeamishness about statistics.

Source: Course concept developed for this guide, informed by Kohavi et al. (2020) and Kramer et al. (2014).

Two smaller obligations close the section, and both are about transparency in directions people forget. The first is internal. An experiment's result is a fact about the business, and the temptation to report only the tests that won is enormous — it is the peeking problem of Section 11.8 relocated from the analysis to the archive. An experimentation practice that keeps a complete log, including the flat results and the losses, is doing the same thing the override log of Section 10.9 did for forecasts: making its own judgment auditable over time, so the organization can learn which of its instincts deserve to keep being funded. The practice that files only winners is not merely dishonest; it is the mechanism by which the retailer in Section 11.14.3's composite vignette built a year of false beliefs.

The second obligation is external, and it is a governance recommendation rather than an empirical claim — a distinction the draft of this chapter blurred and this revision states plainly. It is not universally true that customers cannot be asked to consent to every A/B test, nor that almost no marketing organization discloses experimentation, nor that disclosure has been shown to cost nothing; those are three assertions of fact that would each need a source, and the sources do not exist in the form the claims require. What is true is that consent and disclosure obligations vary, and that the variation follows identifiable factors: the jurisdiction and its privacy regime, the risk the intervention carries, the sensitivity of the outcome being measured, whether the activity is framed as research or as ordinary product improvement, the representations the company has already made in its own privacy and contractual language, and whether the population includes people entitled to additional protection. Some tests in some jurisdictions require explicit consent; many ordinary product-improvement tests do not; and the boundary is a legal and ethical question that an analyst should route to someone qualified rather than resolve from a textbook. Against that background, this guide recommends the plain-language transparency practice as a matter of governance: state, in a privacy or research statement written in language a person would actually read, that the company tests variations of its experiences and messages and that customers may at times receive different versions. The recommendation converts a thing that looks bad when discovered into a thing that was disclosed all along, and an organization that finds itself unwilling to write that sentence has learned something about its own testing program.

11.16 Chapter Summary

This chapter closed Part II by building the only instrument in the guide that produces evidence about actions rather than patterns. It began from an impossibility: a customer's causal effect is the difference between her outcome under treatment and her outcome without it, exactly one of those outcomes ever occurs, and the other — the counterfactual — is not missing from the file but missing from the world. That fundamental problem redirected the enterprise from measuring effects within units to constructing comparisons across them, and the construction is where the chapter's real content lives.

Selection bias named what goes wrong when the comparison is inherited rather than constructed. The Drop Alert's spectacular 5.30-point observational difference decomposed, on four cells of StyleCraft's own record, into a 1.80-point adjusted difference and a 3.50-point composition component — the CRM team's targeting competence read back to it as campaign performance. Two qualifications travel with that decomposition and both are load-bearing. The adjusted 1.80 points is a causal effect only under the miniature's designed assumption that engagement is the sole confounder, measured without error, with no unmeasured selection remaining; it sits on the third rung of Section 11.10's ladder, not the fifth. And the direction of selection bias is not a constant. It points upward when high-propensity customers are preferentially treated, as here, and it points the other way for win-back, save, and onboarding programs that deliberately treat the customers least likely to convert.

Random assignment was then introduced as the device that removes the problem wholesale rather than variable by variable, balancing in expectation every characteristic a customer has, including the ones nobody measured or imagined, at the price of deciding by chance what a competent marketer would rather decide by judgment. The chapter was careful about what that purchase includes. Randomization delivers comparability; an unbiased effect estimate additionally requires that assignment was implemented as planned, that outcomes are observed consistently, that units are analyzed as assigned, that no unit's outcome depends on another's assignment, and that the treatment is well defined at a stated dose. StyleCraft's design was specified as complete randomization — exactly five thousand of ten thousand eligible customers drawn for treatment — which is what makes a realized arm-size shortfall a finding rather than noise.

Around that single move the chapter assembled the protections without which it does not survive contact with an organization. The control arm was defined as the real alternative rather than silence. The unit of randomization was chosen large enough to contain the treatment's spillover, with the reminder that outcomes recorded at a finer grain than the assignment are more description rather than more randomized evidence. The preregistration fixed metric, arms, dose, mechanism, sample size, duration, decision rule, and above all stopping rule before any outcome was visible, as the experimental form of Section 2.5's analytic specification — and every guardrail it declared was one the certified data can actually compute, because a preregistered guardrail that cannot be calculated is a design defect rather than a missing nicety. The power calculation turned a minimum detectable effect into a sample size, and Code 11.7 made it executable: 4,379 customers per arm for the registered 1.5-point effect, 94.5 percent power against the planted 1.8 points at the registered 5,000, and 35.6 percent at a thousand — the number that turns "no significant difference" into a statement about the experiment. The intention-to-treat rule kept every randomized customer in her assigned arm, with an assignment-flow record reported from randomization through outcome observation, because dropping the customers whose messages failed is precisely how an analyst converts a randomized comparison back into an observational one. Covariate balance was reframed as a diagnostic rather than a test, since chance imbalance occurs under a valid procedure and a conspicuous imbalance calls for an audit of the assignment code rather than a re-randomization. And the lab's sham-split procedure was renamed for what it is — an analysis-path check on completed control-arm outcomes — with a genuine end-to-end A/A test named as the separate obligation it remains.

The measurement itself was almost anticlimactic: two proportions, an absolute difference of 1.80 percentage points, a 95 percent interval from 0.81 to 2.79, a relative lift of 30 percent whose own interval — computed as a ratio rather than by rescaling the difference interval — runs from roughly 12 to 50 percent. The chapter's verification theme paid off exactly there, with the estimate recovering the true lift the dataset was built to hide in plain sight. The interval was then converted into a decision by a line the data cannot supply, and the line required a fact the draft of this chapter had left unstated: the dose. At one alert per treated customer, twelve cents against a twenty-eight-dollar contribution gives a break-even lift of 0.43 points, the whole interval clears

it, and the recommendation is robust rather than hopeful. At two alerts the line moves to 0.86 points, the interval's lower edge no longer clears it, and the same evidence supports a different recommendation — which is why the dose belongs beside the metric in the registration, and why the break-even comparison ships with a sensitivity grid across cost and contribution instead of a single number.

Five failure modes were cataloged as the standing ways an honest design is destroyed after the fact — peeking, multiple comparisons, mid-test changes, novelty, and contamination — and the labs made the first tangible by manufacturing false winners out of StyleCraft's control arm and nothing else, separating the roughly 22 percent chance of seeing a significant difference in either direction across fourteen looks from the roughly 11 percent chance of seeing one with the treatment ahead, which is the policy an actual analyst follows. Novelty was given its honest treatment: a comparison of customers assigned in the first and second halves of the window is a temporal-stability diagnostic, not a novelty test, because it compares different cohorts exposed to different drops, and the single binary outcome the certified file carries contains no repeated-exposure information at all. Measuring novelty requires following one randomized assignment across several exposures, which the rollout plan must commission.

A bounded section handled the decisions no chance mechanism can reach, with pre/post-plus-comparison and difference-in-differences taught as arithmetic and their parallel-trends assumption stated as the unverifiable claim it is. And the evidence hierarchy finally placed all of Part II on a ladder, not to rank the chapters' usefulness but to constrain their verbs: descriptions describe, models predict, quasi-experiments suggest under assumptions, and only experiments cause — within a tested population, a tested window, and a tested dose, every one of which the retelling will try to drop. The hierarchy carries one condition the chapter states rather than assumes: it describes the warrant available from competently executed designs, and a randomized experiment wrecked by attrition, noncompliance, interference, outcome switching, or repeated peeking can deliver less than a careful quasi-experiment. The rung is set by the design. The warrant is set by the design and the execution together.

Looking ahead, Part II is now complete, and its output is a shelf of evidence: a segmentation, a driver model, a scored list, a churn program with a threshold, a forecast with an interval, and an experimentally measured lift with a rollout rule. Every one of those artifacts is currently a table, a metric, or a paragraph, and every one of them has to survive a room. That is the problem Part III takes up. The next chapter turns to data visualization for analysis, where the first job of a chart is not to persuade anyone of anything but to let the analyst see what the numbers contain — and where the guide's verification discipline meets a new adversary, the AI-drafted chart that is fluent, attractive, and subtly wrong about the very expansion story these chapters have spent eleven chapters establishing. Between here and there sits Project #1, in which the groups assemble an end-to-end analysis from Part II's machinery and defend it out loud; the exercises below are written with that deliverable in view.

11.17 Exercises for Practice and Homework

The following exercises practice the chapter's main habits: ask how each side of a comparison came to exist before quoting the comparison; declare the metric, the dose, and the stopping rule before the data arrive; size the test against the smallest effect worth finding; audit the assignment and keep every randomized customer in her assigned arm; report absolute lift with its baseline and its interval, and give a relative figure an interval of its own; compare the whole interval against a break-even line computed in advance from a stated dose; and keep the verbs on the rung the design earned. They are organized into three groups. Core chapter practice is the required path and should be completed by every student, and it holds the two homework submissions from which your instructor will assign a subset; Self-Assessment #6 draws on the concept and interpretation material of this chapter. In-class activities are prepared for discussion rather than submitted. Extensions are optional. Each exercise carries its assignment label so that instructors can assign selectively. Note that several of these exercises feed Project #1 directly — groups pursuing the causal track will submit a version of Exercise 11.5 as part of their deliverable, and the rubric in Appendix F applies.

11.17.1 Core Chapter Practice

Exercise 11.1 Concept Check (Required Practice)

Answer each in two or three sentences, in your own words.

- State the fundamental problem of causal inference, and explain why it is not solved by collecting more data about the same customers.
- Define selection bias, explain why competent marketing targeting reliably produces it, and name two marketing programs whose targeting rule would bias an observational estimate downward rather than upward.
- Explain what random assignment guarantees about characteristics the company has never measured, and then name three of the five additional conditions an unbiased effect estimate requires beyond comparability.
- Distinguish Bernoulli randomization from complete randomization, and explain why the distinction determines whether an arm-size shortfall is a finding or a coincidence.
- State the intention-to-treat principle. Then explain, in one sentence, why dropping the customers whose messages failed to deliver converts a randomized comparison into an observational one.
- A colleague sees a 1.4 percent difference in mean prior spend between arms and proposes to re-randomize. Explain why this is a mistake, what the imbalance does and does not tell her, and what the two legitimate responses are.
- Explain why a sham split of completed control-arm outcomes is not an A/A test, and list three parts of the apparatus it does not grade.
- A team randomizes by session but reports conversion per customer. Name the failure and state what it does to the interval's width.

- Distinguish absolute from relative lift, explain which one scales to business volume, and state the assumption under which a relative lift can be carried to a population with a different baseline rate.
- Explain why a confidence interval for a relative lift may not be obtained by dividing the endpoints of a difference interval by the observed control rate.
- Explain the difference between statistical and practical significance, and state the two quantities that must both be declared before a break-even lift can be computed at all.
- Explain why an analyst who checks results daily and stops on significance has a false-positive rate far above 5 percent, and explain why the rate for an analyst who stops only when her treatment is ahead is roughly half the rate for one who stops on any significant difference.
- A test returns no significant difference. Write the honest one-sentence report, including the element Section 11.13 requires that most reports omit.
- Explain why comparing the lift among customers assigned in the first half of an experiment's window with the lift among those assigned in the second half does not test for a novelty effect, and describe the design that would.
- State the parallel trends assumption of difference-in-differences, and explain why showing several parallel pre-periods supports it but does not prove it.
- Place these four claims on the evidence hierarchy and rewrite any whose verb outruns its rung: "VIP members spend 40 percent more"; "the model predicts churn with an AUC of 0.78"; "the store opening caused a $5 per-customer lift relative to comparison markets"; "the early-access alert lifted conversion 1.8 points."
- The evidence hierarchy puts randomized experiments above quasi-experiments. Describe a randomized experiment whose evidence you would trust less than a well-executed difference-in-differences, and name the execution failures that would put it there.

Exercise 11.2 The Observational Trap by Hand (Required Practice)

Using only the four-cell record of Table 11.1 and no code, rebuild every figure in Code 11.2's Verification Check: the four stratum rates, the two within-stratum differences, the two aggregate rates, the naive 5.30-point difference, the constructed counterfactual of 8.25 percent, and the split of 5.30 into a 1.80-point adjusted difference and a 3.50-point composition component. Then show the composition component a second way from mix alone, using the 50-point difference in engagement mix and the 7.00-point untreated gap between engaged and unengaged customers.

Next, change one number and re-derive. Suppose the CRM team had sent the Drop Alert to 5,000 engaged and 3,000 unengaged customers instead of 6,000 and 2,000, with all four stratum rates unchanged. Recompute the treated aggregate rate, the naive difference, and the composition component, and state in one sentence what your result shows about the relationship between targeting precision and the size of the bias.

Then answer the question the chapter cares about most. You have just recovered a 1.80-point adjusted difference from observational data, and the experiment will later recover a 1.80-point effect. Write two sentences explaining why these are not the same finding, naming the

assumptions the first one rests on and the rung each occupies on Table 11.6. Finally, write the two-sentence memo you would send the finance partner, using no statistical vocabulary at all.

Exercise 11.3 Lift, Intervals, and the Break-Even Line (Required Practice)

A different StyleCraft program — a post-purchase SMS asking for a review, hypothesized to raise repeat purchase — was tested on 4,000 customers per arm at a dose of one message per assigned customer. Control: 240 repeat purchasers. Treatment: 296 repeat purchasers. The message costs $0.09 per customer and StyleCraft's average order contributes $28.

Compute the control and treatment rates, the absolute lift in percentage points, and the relative lift. Compute the standard error of the difference and the approximate 95 percent confidence interval on the difference; two standard errors either side is sufficient at this level. Then compute an interval for the relative lift the correct way, as a ratio, using the log construction of Code 11.9, and compare it against the interval the rescaling shortcut would have produced — report both and state which one the deliverable carries. Compute the break-even lift at the stated dose and state whether the decision rule of Table 11.3 is met. Then write the recommendation sentence, including the dose — and note before you start that the pattern from Lab 11.2 Part E may not be the one this result calls for, because a program whose interval does not clear its break-even line needs the honest "the evidence does not settle it" sentence of Section 11.7 rather than the confident one.

Then answer three follow-ups. First: the interval's lower bound and the break-even line are close — describe what a decision-maker should be told, and what additional information would settle it. Second: recompute the break-even line at a contribution of $22 rather than $28, state which way the line moves and whether your conclusion changes, and then say which of the two contribution figures you would put in the deliverable and why the other one belongs there too. Third: the CRM lead proposes running the test another two weeks "to tighten the interval." Explain, citing the relevant section, why this is or is not acceptable, and what would make it acceptable.

Exercise 11.4 Design the Test (Required Practice)

For each scenario, produce a complete preregistration using every row of Table 11.3 — including the treatment-dose row, the assignment-mechanism row, and the instrumentation obligation row — and state the unit of randomization with one sentence of justification and one sentence naming the contamination path you are guarding against. For every guardrail you declare, name the field it would be computed from; if that field does not exist, say so and move the guardrail to the instrumentation obligation rather than declaring something you cannot measure.

- StyleCraft wants to know whether adding a free-returns message to the checkout page raises order completion.
- StyleCraft wants to know whether a new in-store styling appointment service raises revenue per visit, and can staff it in at most four stores.

- StyleCraft wants to know whether its paid social spend produces incremental revenue at all, in a market where the platform cannot suppress ads for chosen individuals.
- StyleCraft wants to know whether raising the loyalty program's points-per-dollar rate for Insider members increases their purchase frequency.
- StyleCraft wants to know whether the Drop Alert's effect decays with repeated exposure — the novelty question Section 11.8 says the certified file cannot answer. Specify the design fully: how many drops the assignment is held across, what outcome is recorded after each exposure, what quantity estimates the decay, and what sample size each period-specific comparison would need at 80 percent power to detect a half-point change in the lift.

For the second and third scenarios, state whether a randomized experiment is feasible; where it is not, specify the Section 11.9 design you would use instead, name the comparison group, and state the assumption on which the estimate would rest and the pre-period evidence you would show to support it.

Exercise 11.5 The Full Experiment Deliverable (Homework Submission)

Using the certified ab_test file, produce the complete deliverable described in Section 11.13, as a Colab notebook plus a two-page written report. The notebook must contain, in order and using the chapter's cell numbering as its outline: the data contract of Code 11.3, the assignment-flow table of Code 11.4, the covariate balance profile of Code 11.5 with missing values represented explicitly, the sham-split analysis-path check of Code 11.6, the sample-size and power calculation of Code 11.7 including its sensitivity table, the registered intention-to-treat analysis of Code 11.8, both intervals from Code 11.9 with the rescaling shortcut shown and labeled, the peeking simulation of Code 11.10 reporting all three stopping policies, the single-experiment p-value path of Code 11.11, the break-even comparison of Code 11.12 including its dose-and-contribution sensitivity grid, and the temporal-stability diagnostic of Code 11.13 with its standard errors.

The report must contain the registration with its dose row, the assignment certificate, the primary result in percentage points and in dollars, the relative figure with its own interval, the decision with its rule and its sensitivities, the limitations naming population, window, dose, novelty, and the unmeasured paid social counterfactual, and the guardrail and reversal plan with the instrumentation obligation named. Include a comparison against the known true lift published in the data dictionary's design note, and state in one sentence what that comparison does and does not verify. State explicitly, in one sentence each, which of your exhibits is an A/A test and which is not, and what the multiple of the observational headline over the experimental result actually is — against the point estimate and against the interval's upper bound, which are different numbers. Groups on Project #1's causal track may submit an extended version of this deliverable; see Appendix F for the rubric.

Exercise 11.6 AI Experiment Audit (Homework Submission)

Run the four-move round trip of Lab 11.2 Part E and document it per Appendix D. Submit: (a) your registration prompt and the assistant's draft, with the UNSPECIFIED rows counted, a note on whether it invented a treatment dose or left it unspecified, and a paragraph on what that count revealed about your own prompt; (b) the deliberately misleading analysis prompt and the assistant's response, audited line by line against Table 8.5's five points and Table 11.7's three, with each failure you found named and quoted — check specifically for a per-protocol drift that silently dropped undelivered customers and for a relative-lift interval built by rescaling a difference interval; (c) the adversarial prompt's output, with the items you kept and the items you discarded, and one sentence of justification for each discard; (d) the assistant's drafted recommendation paragraph, your marked-up version showing every verb you changed and why, and your final version, which must name the tested dose. Your grade rests on the audit, not on the assistant's output: an exchange in which the assistant performed badly and you caught everything is a better submission than one in which it performed well.

11.17.2 In-Class Activities

Exercise 11.7 Spot the Testing Mistake (In-Class Discussion)

Each of the following is a real pattern in commercial testing. Name the failure from Table 11.5, state the mechanism in one sentence, describe what the report would look like, and state the single change that would have prevented it. Two of the items contain a failure Table 11.5 does not list; name those from elsewhere in the chapter.

- "We launched Monday and by Wednesday the treatment was up 11 percent with $p = 0.03$, so we rolled it out Thursday."
- "Conversion was flat overall, but revenue per customer was up 6 percent and significant, so we're calling it a win."
- "Halfway through we noticed the treatment creative had a typo, so we fixed it and let the test keep running."
- "The new referral offer lifted signups 22 percent in the test. After three months of full rollout, signups are back where they started."
- "The 15-percent-off code we tested with 5,000 customers ended up on a coupon aggregator site in week one, but we kept the test running."
- "We ran the test on our email-engaged segment because they respond fastest, and we're rolling the winner out to the full file."
- "About 8 percent of the treatment arm's messages bounced, so we removed those customers and compared the rest against the full control arm."
- "The interval on the lift was 0.4 to 3.1 points and our break-even is 1.2 points, but the point estimate is 1.75, so we're going ahead."

Exercise 11.8 The Holdout Meeting (In-Class Discussion)

StyleCraft's VP of Marketing proposes a twelve-month holdout: 5 percent of the customer base, chosen at random, will receive no marketing communications at all for a year, so that the total incremental value of the CRM program can finally be measured. The CRM manager objects that the company would be knowingly withholding offers, early access, and loyalty communications from four hundred real customers for a year. The finance partner points out that nobody currently knows whether the CRM program creates value at all, and that the company spends several million dollars a year on it.

Prepare a position. Evaluate the proposal against all four conditions of the "Holdout Is a Debt You Repay" concept box, and state which conditions it meets, which it fails, and how it could be redesigned to meet them all while still answering the question. Address explicitly: how large the holdout actually needs to be given the effect size worth detecting, computed rather than asserted, using the machinery of Code 11.7; whether twelve months is required or merely convenient; what the customers in the holdout should be told, if anything, and what governs that answer rather than what your instinct says; what repayment would look like at the end; and whether your answer would change if the holdout were selected by a rule — say, lowest-value customers — rather than at random, and what that change would do to both the ethics and the estimate. Then address one further question the chapter raises: this holdout receives *no* communications rather than the existing experience, which is a departure from the bounded-harm condition. State whether that departure is defensible here, and what would have to be true for it to be.

11.17.3 Extensions

Exercise 11.9 Peeking Under Two Stopping Rules (Optional)

Code 11.10 reports three false-positive rates for the same simulated experiments: the final-look rate, the any-significant-difference rate, and the treatment-ahead rate. Extend the simulation in three directions and report what changes. First, vary the number of looks from 2 to 28 and plot all three rates against the number of looks; describe the shape of each curve and explain why the any-direction curve flattens rather than approaching the independent-look bound of $1 - 0.95$ raised to the number of looks. Second, vary the minimum arm size required before a look is taken, from 50 to 1,000, and explain the direction of the effect on the inflation. Third, and this is the one that changes how you read commercial test reports, add a fourth policy: an analyst who stops when the treatment is ahead *and* the observed lift exceeds one percentage point, which is what a business rule with a practical-significance threshold actually looks like. Report its false-positive rate, compare it against the other three, and write one paragraph on whether adding a practical-significance threshold to an undisciplined stopping rule fixes the problem, reduces it, or merely hides it. State a seed for every figure you report, and report the spread across at least five seeds rather than a single decimal.

11.18 Glossary of Terms

This glossary includes only the terms introduced in this chapter. Each definition is tied to the sources used in the chapter rather than added for decoration.

A/A test. An experiment in which both arms receive the identical experience, run through the full production assignment, delivery, and analysis pipeline, so that the known-zero true effect grades the machinery rather than any intervention. Because the nominal false-positive rate is not zero, a single significant result is an occasion for investigation rather than a verdict; repeated or systematic failures indicate a defect in randomization, delivery, or measurement (adapted from Kohavi et al., 2013).

Absolute and relative lift. Absolute lift is the treatment-minus-control difference in the metric's own units, in percentage points for a conversion rate, and is the quantity that scales to business volume. Relative lift is that difference as a proportion of the control rate; it is systematically the larger and more persuasive figure whenever the baseline is small, and it transports to a population with a different baseline only under a stated multiplicative-effect assumption. Each requires its own confidence interval, the second constructed as a ratio (adapted from Kohavi et al., 2020).

Break-even lift. The lift at which a program exactly pays for itself: the cost per assigned customer divided by the contribution per incremental conversion. It is undefined until the treatment dose fixes the cost per assigned customer, it belongs in the preregistration's decision rule, and it is compared against the whole confidence interval rather than the point estimate, with its sensitivity to cost and contribution reported beside it (course concept developed for this guide, informed by Kohavi et al., 2020).

Causal inference. The estimation of what an outcome would have been under an alternative exposure, using observed data plus design features or assumptions that make an observed group a credible stand-in for the unobserved counterfactual; its credibility rests on research design rather than on dataset size or model sophistication (adapted from Holland, 1986; Imbens & Rubin, 2015).

Contamination. Influence of one unit's treatment on another unit's outcome — also called interference or spillover — through shared offers, shared inventory, shared households, or market-level effects; it biases the measured difference and is addressed by enlarging the unit of randomization rather than by post hoc adjustment (adapted from Kohavi et al., 2020).

Counterfactual. The potential outcome that did not occur — the result the unit would have shown under the exposure it did not receive; permanently unobservable, and therefore estimated by comparison across units (adapted from Rubin, 1974; Holland, 1986).

Difference-in-differences. A quasi-experimental estimator that compares the change in outcomes over time in a treated group with the change over the same period in an untreated comparison group, taking the difference of the two changes; valid only under the

unverifiable parallel trends assumption, which is supported but not proven by matching pre-period trends (adapted from Card & Krueger, 1994; Angrist & Pischke, 2015).

Evidence hierarchy. An ordering of evidence by the strength of causal claim it supports — anecdote, description and association, model-based association and prediction, quasi-experiment, randomized experiment — whose practical function is to hold a conclusion's language to the strength of the design that produced it. It describes the warrant available from competently executed designs; execution failures can move a study's real warrant below its rung (adapted from Angrist & Pischke, 2015; Kohavi et al., 2020).

Intention-to-treat analysis. Comparison of units according to the condition they were randomly assigned, regardless of whether the treatment was delivered, received, or complied with. It preserves the comparability random assignment created, because groups defined by post-assignment events are no longer randomized; it estimates the effect of being offered the program, which is a different quantity from the effect of the treatment actually received (adapted from Hopewell et al., 2025; Imbens & Rubin, 2015).

Multiple comparisons. The inflation of false-positive risk that occurs when many metrics or segments are tested and the significant ones are reported as though each had been the only test; remedied by declaring one primary outcome in advance and treating unplanned findings as hypotheses for a future experiment (adapted from Simmons et al., 2011; Kohavi et al., 2020).

Novelty effect. A response to a treatment arising from its unfamiliarity rather than its enduring value, producing an initial effect that decays with habituation; its mirror image, primacy, is initial underperformance from disrupted habit. Because both concern change across repeated exposures, detecting either requires outcomes observed over several exposures for the same randomized units rather than a comparison of cohorts assigned at different dates (adapted from Kohavi et al., 2020).

Peeking. Examining an experiment's results before the declared sample size or duration is reached and allowing what is seen to influence stopping or analysis; because each look offers a fresh chance to observe a threshold-crossing fluctuation, it raises the false-positive rate far above the nominal level, undetectably in the final output. The inflation is larger when hits in either direction count and roughly half as large under the realistic policy of stopping only when the treatment is ahead (adapted from Johari et al., 2017).

Potential outcomes. The pair of results a unit would show under treatment and under no treatment, whose difference is the unit's causal effect; exactly one is ever observed, which is the fundamental problem of causal inference (adapted from Rubin, 1974; Holland, 1986).

Preregistration. A dated, unalterable record of an experiment's hypothesis, primary metric, arms, treatment dose, assignment mechanism, unit of randomization, eligible population, sample size, duration, analysis, decision rule, and stopping rule, created before the

outcomes are observed; it is what distinguishes a tested prediction from a description of the data that generated it. Every guardrail it declares must be computable from the data the pipeline actually delivers (adapted from Nosek et al., 2018).

Random assignment. Allocation of units to experimental conditions by a chance mechanism independent of every characteristic of the units, making the arms comparable in expectation on all characteristics, measured or not. Under Bernoulli randomization each unit is assigned independently and arm sizes are random; under complete randomization a fixed number of units is drawn and arm sizes are exact. Comparability yields an unbiased effect estimate only when assignment is implemented as planned, outcomes are observed consistently, units are analyzed as assigned, interference is absent, and the treatment is well defined at a stated dose (adapted from Fisher, 1935; Imbens & Rubin, 2015).

Selection bias. The difference between an observed group comparison and the causal effect it is taken to estimate, arising because treated and untreated units differ systematically in ways that affect the outcome; a property of how the groups came to exist, unaffected by dataset size, and directed upward or downward according to the selection mechanism rather than always upward (adapted from Angrist & Pischke, 2009; Imbens & Rubin, 2015).

Sham-split analysis-path check. This guide's name for a diagnostic that divides completed control-arm outcomes at random into two groups and runs the primary test on them. Because the true effect is zero by construction, it grades the metric definition, the test statistic, and the analysis code, and it establishes the nominal false-positive behavior of that path. It grades neither the production randomizer nor delivery nor tracking nor assignment-to-exposure integrity, and it is therefore not an A/A test (course concept developed for this guide, informed by Kohavi et al., 2013).

Statistical power. The probability that an experiment detects an effect of a specified size at a specified significance level, given sample size and outcome variability; required sample size grows roughly with the inverse square of the smallest effect worth detecting, and underpowered tests overstate the effects they do declare significant (adapted from Cohen, 1988; Kohavi et al., 2020).

Treatment and control groups. The arm receiving the intervention under test and the arm receiving the existing experience or declared alternative; the control supplies the estimate of what the treatment arm would have shown absent the treatment, and should represent the decision's real alternative rather than a strawman (adapted from Fisher, 1935; Imbens & Rubin, 2015).

Treatment dose. The number and timing of exposures a treated unit receives, stated as part of the intervention's definition. The dose fixes the cost per assigned customer and therefore the break-even lift; an experiment run at one dose licenses a rollout at that dose, and a rollout at a different frequency is a different intervention whose economics and whose response

must both be re-derived (course concept developed for this guide, informed by Kohavi et al., 2020).

Unit of randomization. The entity to which the chance mechanism assigns a condition — customer, session, store, market, or time period — which fixes the effective sample size and the grain the analysis must respect. Outcomes may be recorded more finely, but treatment variation exists only at the assignment grain, and the unit should be at least as large as the unit through which the treatment can reach a person (adapted from Kohavi et al., 2020).

11.19 Further Readings

Students who want additional background may begin with the following readings. The applied treatments are listed first, evidence and methods second, reporting standards last.

- Kohavi et al. (2020) for the standard practitioner's book on controlled experiments — written from a decade of platform-scale practice, and the direct expansion of Sections 11.4 through 11.8. The chapters on metrics, common pitfalls, and organizational maturity are the most useful pages a new marketing analyst can read on this subject, and the treatment of guardrail metrics is where the instrumentation obligation of Table 11.3 comes from.
- Angrist and Pischke (2015) for a genuinely accessible introduction to the logic of causal design, including randomized trials and difference-in-differences — the level above this chapter's Sections 11.2, 11.3, and 11.9 without the econometric prerequisites of their earlier volume.
- Johari et al. (2017) for the definitive practical treatment of peeking — why continuous monitoring corrupts conventional tests, how large the corruption is, and what valid sequential alternatives look like; the paper behind Code 11.10.
- Gordon et al. (2019) for the head-to-head comparison of experimental and observational advertising measurement at scale, and the clearest available evidence for why Section 11.3's warning about statistical adjustment is not a theoretical worry — including the finding that the divergence runs in both directions rather than consistently upward.
- Lewis and Rao (2015) for the power arithmetic of advertising measurement — the argument that many reported advertising returns come from tests that could not have measured them, which is Code 11.7's sensitivity table carried to its industry conclusion.
- Nosek et al. (2018) for preregistration as a general scientific reform, and for the distinction between prediction and postdiction that Section 11.5 borrows wholesale.
- Hopewell et al. (2025) for the CONSORT 2025 reporting guideline, which is where the flow record of Table 11.4 and the intention-to-treat discipline of Section 11.5 come from in their mature form. A marketing analyst who reads the participant-flow and analysis-population items once will never again drop a randomized customer without noticing.

- Senn (1994) and Altman (1985) for the argument that baseline significance tests are the wrong instrument for judging whether a randomization worked — short, pointed, and the direct source of Section 11.5's reframing of the balance check as a diagnostic.

11.20 References

Altman, D. G. (1985). Comparability of randomised groups. *Journal of the Royal Statistical Society: Series D (The Statistician), 34*(1), 125–136. https://doi.org/10.2307/2987510

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Angrist, J. D., & Pischke, J.-S. (2015). *Mastering 'metrics: The path from cause to effect*. Princeton University Press.

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives, 31*(2), 3–32. https://doi.org/10.1257/jep.31.2.3

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

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review, 84*(4), 772–793.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

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

Hopewell, S., Chan, A.-W., Collins, G. S., Hróbjartsson, A., Moher, D., Schulz, K. F., Tunn, R., Aggarwal, R., Berkwits, M., Berlin, J. A., Bhandari, N., Butcher, N. J., Campbell, M. K., Chidebe, R. C. W., Elbourne, D., Farmer, A., Fergusson, D. A., Golub, R. M., Goodman, S. N., … Boutron, I. (2025). CONSORT 2025 explanation and elaboration: Updated guideline for reporting randomised trials. *BMJ, 389*, e081124. https://doi.org/10.1136/bmj-2024-081124

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.

Johari, R., Koomen, P., Pekelis, L., & Walsh, D. (2017). Peeking at A/B tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1517–1525). https://doi.org/10.1145/3097983.3097992

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013). Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1168–1176). https://doi.org/10.1145/2487575.2488217

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

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, 111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Lewis, R. A., & Rao, J. M. (2015). The unfavorable economics of measuring the returns to advertising. *Quarterly Journal of Economics, 130*(4), 1941–1973. https://doi.org/10.1093/qje/qjv023

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701. https://doi.org/10.1037/h0037350

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference* (pp. 92–96). https://doi.org/10.25080/Majora-92bf1922-011

Senn, S. (1994). Testing for baseline balance in clinical trials. *Statistics in Medicine, 13*(17), 1715–1726. https://doi.org/10.1002/sim.4780131703

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

statsmodels developers. (2026a). *statsmodels.stats.proportion.confint_proportions_2indep* [Software documentation]. https://www.statsmodels.org/stable/generated/statsmodels.stats.proportion.confint_proportions_2indep.html

statsmodels developers. (2026b). *Statistics: statsmodels.stats* [Software documentation]. https://www.statsmodels.org/stable/stats.html

Vaver, J., & Koehler, J. (2011). *Measuring ad effectiveness using geo experiments*. Google Inc.

Verma, I. M. (2014). Editorial expression of concern: Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, 111*(29), 10779. https://doi.org/10.1073/pnas.1412469111

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108