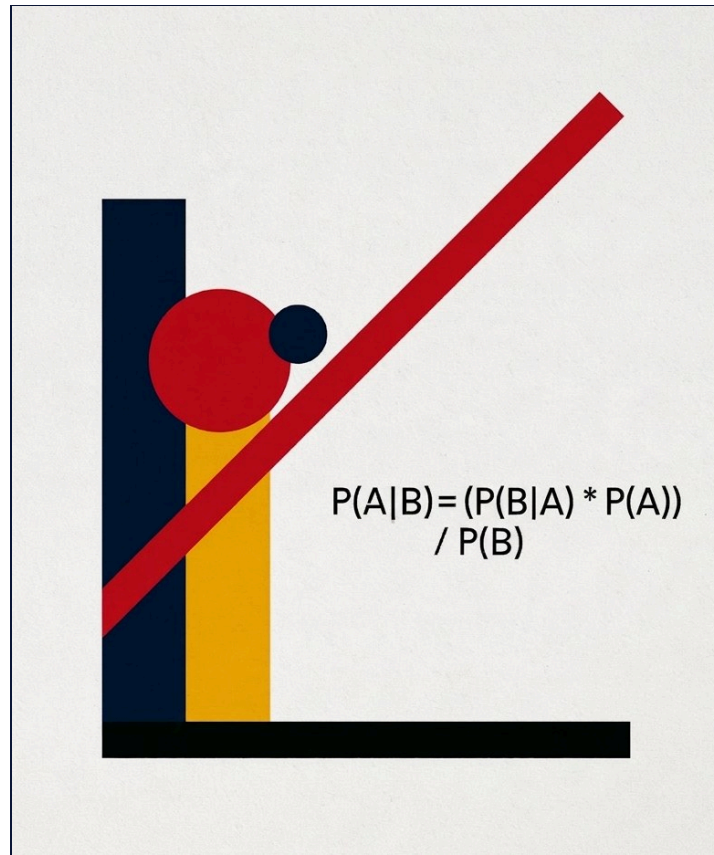




BAYESIAN MARKETING

*How to Make Better Marketing Decisions with Statistics
and AI*

ATMA DEGEYNDT

A NOTE BEFORE WE BEGIN

I've sat through enough marketing meetings to recognize the moment when a number becomes a story.

An ad generates a handful of sales. Someone says the message is working. Another has a different explanation: the audience was unusually receptive, the platform finally figured out delivery, or a competitor went quiet. The debate usually grows more confident as the evidence gets thinner.

A dashboard records what happened under its measurement rules. It doesn't automatically tell us why it happened, how reliably it will happen again, or what we should do with the next \$10,000.

This book is about closing those gaps.

Bayesian statistics gives us a disciplined way to combine what we thought before an experiment with the evidence that arrives afterward. Instead of concealing uncertainty behind one average, it lets us examine which outcomes remain plausible and how those probabilities change. The method is particularly useful when observations arrive over time and decisions cannot wait for perfect information.

There is no mathematical exemption from bad judgment. A model built on irrelevant data can be confidently wrong. A carefully computed conversion probability does not establish that advertising caused the sale. And a 99% probability of a higher click rate is no substitute for understanding the economics of acquiring customers.

We'll follow one fictional company through the book. North Pier sells an \$80 insulated lunch bag. It runs two messages, makes ordinary marketing mistakes, and gradually learns to ask better questions. Every North Pier number is invented for instruction and stays fixed unless a new scenario is explicitly identified.

The exercises start with arithmetic you can verify on paper. They progress through probability distributions, decisions under uncertainty, experiment design, connected funnel models, and causal measurement. Then we'll bring in AI, which can now produce statistical code and analysis quickly enough to make competent supervision more valuable than ever.

No programming is required to follow the argument. The optional code and AI exercises are there for readers who want to practice. Nothing depends on proprietary software or on believing that the author's system is finished merely because a technical specification exists.

One habit runs through everything that follows. Before treating a result as an answer, decide exactly which question it answered.

CONTENTS

PART I WHAT AN OBSERVATION IS WORTH



CHAPTER ONE The Ad That Got Judged Too Soon



CHAPTER TWO What You Believed Before the Click



CHAPTER THREE Small Numbers, Loud Opinions

PART II FROM STATISTICS TO A DECISION



CHAPTER FOUR The Whole Distribution, Not the Average



CHAPTER FIVE Priors You Can Defend



CHAPTER SIX Buy an Answer Worth Knowing



CHAPTER SEVEN Stop, Continue, Hold, or Switch

PART III THE BUSINESS BEHIND THE DASHBOARD



CHAPTER EIGHT One Funnel, Several Linked Questions



CHAPTER NINE Reported Is Not Caused



CHAPTER TEN When Yesterday Stops Describing Today

PART IV

AI AND PROFESSIONAL JUDGMENT



CHAPTER ELEVEN An AI Analyst with a Supervisor



CHAPTER TWELVE Build a Practice That Can Say “We Don’t Know”

APPENDICES

APPLY THE METHODS



APPENDIX A The North Pier Workbook



APPENDIX B A Reproducible First Bayesian Analysis



APPENDIX C A Decision Memo You Can Reuse



APPENDIX D Essential Terms Without the Mystique



APPENDIX E A Complete Decision Review, Not Just a Statistical Report



APPENDIX F The AI Prompt Lab: Four Tasks Worth Delegating



APPENDIX G A Design Case: The Bayesian Marketing Decision Framework

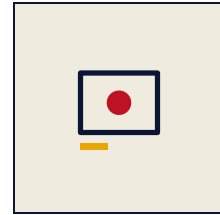


PART I

WHAT AN OBSERVATION IS WORTH



THE AD THAT GOT JUDGED TOO SOON



A founder shows an ad to a friend.

“I don’t like it,” the friend says.

The next morning, the founder wants to change the campaign. The headline, the image, maybe the whole audience. None of the proposed changes comes from customer behavior. There has been no new experiment, no purchase data, not even a useful account of what the friend disliked. One person reacted, and a decision was born.

I’ve encountered some version of this exchange more times than I care to count. It happens with ad creative, websites, pitch decks, product names, and pricing. It’s tempting because the feedback feels like evidence.⁴ Someone has given us a reaction we can remember. We can hear their voice. We know whether they seemed confident. There is nothing abstract about it.

But how much does it tell us?

The friend may be right. That’s a possibility worth taking seriously. The question is whether the reaction tells us anything reliable about the people we’re trying to reach. Is this friend a likely customer? Have they been asked to buy the product or merely asked for an opinion? Is the message unclear, or is it simply not to their taste? Those aren’t evaluations. They determine what evidence, if any, we’ve received.

Most marketing mistakes don’t begin with someone refusing to look at the numbers. They begin when we assign the wrong weight to something we’ve observed.

For a marketer, that problem gets expensive quickly.

■ TWO ADS, ONE MISTAKEN CONCLUSION

Consider a fictional startup called North Pier. It sells a durable insulated lunch bag for \$80. The company is trying two ads. One says, “Pack smarter.” The other says, “Lunch stays cold.”

The first ad, A, makes an appeal about organization. The second, B, names a practical benefit. Both ads begin running. After two days, the numbers look like this:

MEASURE	AD A	AD B
Impressions	2,000	2,000
Clicks	40	34
Tracked orders	1	2
Spend	\$60	\$60

A click-through rate, or CTR, is clicks divided by impressions. Ad A has a 2.0% CTR. B has 1.7%. A appears stronger on the metric at the top of the advertising funnel.

But B has two tracked orders and A has one. So is B better?

Not so fast. We have three orders in total. One additional order has doubled B’s result relative to A. That is a dramatic percentage difference built on almost no sales data.

And we’re still not asking the right question. What does “better” mean? More clicks? More purchases? More profit? A lower cost to acquire customers who actually remain customers?

North Pier has a simple answer to part of that question. Each \$80 sale costs \$48 to produce, fulfill, process, and cover expected variable returns. That leaves \$32 before advertising. A customer acquired for \$40 isn’t profitable on this simplified first-order calculation, even if the ad attracts plenty of clicks. A customer acquired for \$20 is.

After two days, A has spent \$60 to generate one tracked order. B has spent the same amount to generate two. We know what the platform reports. We do not yet know whether those differences are persistent, whether platform delivery exposed the same kinds of people to the ads, or how many of those orders would have happened without advertising.

The honest report is not “A won” or “B won.”

It's this: A has drawn more clicks. B has produced more tracked orders. With so few orders, the uncertainty is substantial. We should not buy a large new budget allocation with this evidence alone.

That's a useful conclusion. It may be less satisfying than a winner, but it is more defensible.

■ NUMBERS DO NOT TELL US HOW MUCH TO BELIEVE THEM

Every advertising dashboard contains a small trap. Its numbers look complete.

Two percent CTR. Three orders. \$60 spent. The interface rarely looks uncertain. If the figures are displayed to one or two decimal places, they can even look precise.

But a rate is an observation, not automatically our best estimate of what will happen next.

Suppose you flip a coin five times and get four heads. The observed rate is 80%. That doesn't give us a good reason to predict an 80% chance of heads on the next flip. Five trials are too few to establish that conclusion, especially when our previous experience tells us ordinary coins are usually close to fair.

An advertising rate isn't a coin toss in every respect. Platforms choose which users see ads; creative can fatigue; people differ. Still, the lesson holds: the amount and quality of the evidence matter. A 2% click rate calculated from 40 clicks should not be treated with the same certainty as a 2% rate calculated from 40,000.

The psychologist Daniel Kahneman and Amos Tversky studied this habit of drawing confident conclusions from samples that are too small to support them. Their work on the "law of small numbers" remains recognizable to anyone who has watched a founder react to an early dashboard.¹

It isn't only founders. Experienced marketers can be just as vulnerable. One lousy afternoon looks like a collapsing campaign. One unusually good conversion looks like proof the new creative works. Someone calls it a trend. Someone else calls it a breakthrough. The data hasn't earned either word.

■ THE NEXT TWELVE DAYS

North Pier doesn't stop measuring. It also doesn't pretend that waiting is always wise. A broken checkout would justify immediate action. An obvious factual error in an advertisement should be corrected immediately. Those are different problems from deciding which message will acquire customers profitably over time.

The ads continue under a bounded test. Fourteen days after launch, the cumulative results are:

MEASURE	AD A	AD B
Impressions	20,000	20,000
Clicks	400	340
Tracked orders	8	18
Spend	\$500	\$500

Ad A still wins on CTR: 2.0% against 1.7%. If the marketing team rewards clicks, it could continue reporting A as the better ad.

But A's 8 tracked orders represent \$640 in revenue. B's 18 represent \$1,440. And each tracked order, under our simplified assumptions, contributes \$32 before advertising.

Eight orders contribute \$256. Subtract \$500 in ad spend: A is down \$244.

Eighteen orders contribute \$576. Subtract \$500: B is up \$76.

That changes our working judgment. It does not prove B's message caused the difference. We haven't established that this was a randomized experiment. The platform may have delivered the messages to different types of customers. We haven't established that every tracked sale is incremental. Later we'll deal with both problems.

But we can now see how expensive it would be to mistake click-through rate for the objective. A produces more clicks and, in this accounting view, loses money. B produces fewer clicks and performs better on tracked first-order contribution.

The stronger decision isn't to worship B. It is to recognize that the evidence now supports a different question: can we reproduce B's apparent commercial advantage under a test that separates the message from the audience and gives us a better estimate of the customers advertising actually brings in?

■ WHAT BAYESIAN THINKING CHANGES

Bayesian statistics gives us a disciplined way to keep our starting knowledge in view while allowing new observations to change our estimates. It asks how strongly an observation favors one explanation over another, given what we knew before.²

That doesn't mean the previous belief was true. It doesn't mean strong beliefs should survive contradictory evidence. And it certainly doesn't mean we can recover more information than our sample contains.

It means an early observation is not automatically a verdict.

The practical questions become more precise. What did we believe before the test? Why? How likely would these results be under different possible explanations? How uncertain are we now? And what would it cost us to act as though we were certain?

There's a second question that marketing teams often skip: are we looking at the right outcome in the first place?

North Pier's numbers give us a preview of both mistakes. Early percentages encouraged a verdict before enough data arrived. Later, CTR continued to favor the ad that performed worse on the business outcome. The problem wasn't a shortage of reports. It was the relationship between what we measured and what we were trying to decide.

Bayesian methods will not answer that second question for us. A beautifully calculated probability about the wrong target can still support a bad decision.

Before doing any math, define the decision.

For North Pier, the decision isn't "Which ad is more popular?" It is "Which approach should receive more money if our goal is to acquire incremental customers at an acceptable cost?"

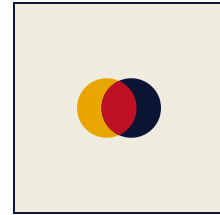
Now we have a question worth testing.

■ TRY IT ON YOUR OWN CAMPAIGN

Find one recent campaign decision. Write down the observation that triggered it and the conclusion someone drew. Separate the two. Then identify the missing information that would most change the conclusion. It might be the number of impressions, the profile of the people sampled, the cost of each sale, or the difference between tracked and incremental orders.

If the answer is “we don’t know yet,” don’t soften it. That’s information your next decision needs.

WHAT YOU BELIEVED BEFORE THE CLICK



Let's return to North Pier on day two. Ad A has 40 clicks from 2,000 impressions. Ad B has 34.

Ad A has the higher observed click-through rate. We know that. But what should we expect from these ads tomorrow, or next week? How sure should we be?

The dashboard gives us two percentages. Bayesian analysis gives us a way to turn those observations into an estimate with uncertainty attached.

The logic begins before the first click.

■ YOUR PRIOR IS ALREADY IN THE ROOM

No marketer arrives at a campaign with a perfectly empty mind. You know something about the product, the platform, the audience, and earlier campaigns. Maybe you've run dozens of similar ads. Maybe you've never advertised this kind of product before.

That previous information matters. Bayesian statistics makes us say what it is and how much influence it deserves.

A prior is a probability distribution representing what the model considers plausible before it sees the data from this particular test. That's not the same as the founder's conviction that their headline is brilliant. A defensible prior comes with a source and assumptions. The wrong prior can mislead us. It should be checked, not protected.³

For this teaching example, suppose we start both ads with the same prior for CTR: a Beta distribution with parameters 2 and 98, written Beta(2,98). Its average is 2%. It is a deliberately light starting point, roughly equivalent in this simple model to 100 prior impression trials, but it is not a recommended default for every campaign. In a real account, we'd examine relevant historical campaigns and how similar they really were.

Why use a Beta distribution? A click either occurs or doesn't occur on an impression opportunity, and CTR lies between 0 and 1. For a simplified binomial model of independent trials with a stable probability, a Beta prior is mathematically convenient. It produces another Beta distribution after new observations arrive.

There is a catch. Real advertising impressions aren't necessarily independent or identical. One person may see the ad repeatedly. Platform auctions select who sees it. The underlying rate may change during the test. The simple model is an entry point, not a complete account of an ad platform.

■ THE RULE BEHIND THE RECIPE

The bookkeeping is simple for a Beta prior and a binomial likelihood: add clicks to alpha and nonclicks to beta. But there is a reason it works, and you need it if you ever leave this special case.

Posterior $\propto$ prior $\times$ likelihood. Read that as: our updated assessment is proportional to what we thought was plausible before the test, multiplied by how well each possible explanation predicts the evidence we actually received. The result is scaled so all the probabilities add up to one. This is Bayes' rule. It is the same mechanism whether the unknown quantity is CTR, order value, or a more complicated collection of parameters.²

Here is a separate, deliberately simple probability example. Suppose our prior probability that a campaign is genuinely effective is 20%. We observe a particular favorable signal. Assume that signal appears 60% of the time when the campaign is effective, but also 20% of the time when it isn't. Then:

$$\text{Posterior probability} = (0.60 \times 0.20) \div [(0.60 \times 0.20) + (0.20 \times 0.80)] = 42.9\%.$$

The signal is three times more likely under the effective-campaign explanation. It raises our estimate from 20% to nearly 43%. It does not make the campaign a certain winner. The example is invented to teach the mechanism, not presented as North Pier data.

A prior is not a mood. A likelihood is not a compliment from an AI assistant. Both are explicit assumptions about what kinds of results we would see. If they're badly specified, the calculation can be perfectly executed and still lead us astray.

■ THEN WE SEE THE DATA

In the simple Beta-binomial model, we update the first parameter with clicks and the second with impressions that did not produce clicks.

Ad A started with $\text{Beta}(2,98)$. It received 40 clicks and 1,960 nonclicks, producing $\text{Beta}(42,2058)$. Ad B received 34 clicks and 1,966 nonclicks, producing $\text{Beta}(36,2064)$.

Those updated distributions are called posteriors. They describe which values of CTR remain plausible under the model after incorporating the observations.

The posterior mean for A is 42 divided by 2,100, or 2.00%. For B it is 36 divided by 2,100, or approximately 1.71%. The raw rates were 2.00% and 1.70%, so the updated averages have not moved much. That isn't surprising: the prior was intentionally weak relative to the 2,000 new impressions.

But a posterior is not just an average. It gives us a range of plausible values and a way to ask questions about comparisons.

Under this simplified model, a central 95% credible interval for A's CTR is approximately 1.45% to 2.64%. B's is approximately 1.20% to 2.31%. These ranges overlap substantially. We should expect some variation in the observed numbers if we repeated the exercise.

This is what the dashboard hid when it displayed 2.0% versus 1.7% as if the difference were a settled fact about future performance.

A credible interval needs precise language. Under the stated model, after observing the data, the interval contains 95% of the posterior probability assigned to the underlying CTR parameter. It is not a promise that 95% of future observed CTRs will fall inside those numbers. Predictions of future campaign measurements have their own uncertainty.²

■ THE PROBABILITY OF BEING BETTER IS NOT THE VALUE OF BEING BETTER

If we draw matched hypothetical CTRs from each ad's posterior, A's draw is higher than B's in roughly three out of four comparisons. Under the simple independent-prior model, the probability A has the higher CTR is about 75%.

That's a useful statement, if we say what assumptions produced it. It is not a declaration that A has a 75% probability of being the more profitable ad. It isn't even a causal claim that A's wording makes people click more. Our delivery setup may have shown the

ads to different audiences.

And suppose A really does have a slightly higher CTR. So what? If the clicks rarely turn into profitable customers, the difference may be nearly worthless.

In a serious decision report, we'd ask two separate questions: how likely is A to have the higher CTR, and what would it be worth to us if it does?

A probability does not contain the business objective. We have to supply that.

■ PRIORS ARE NOT A LICENSE TO BE STUBBORN

One reason priors provoke such strong reactions is that marketers hear “prior belief” and imagine an analyst smuggling personal preference into the model.

That is a real risk. But treating all priors as suspect doesn't eliminate assumptions. It merely makes them harder to find.

Suppose two advertisers test the same creative. One has thousands of closely comparable earlier impressions. The other has no relevant history. Starting both from the same generic distribution would throw away useful information in the first case. On the other hand, treating last year's performance in a different country as though it were identical evidence would create false confidence.

A good prior requires more than finding a benchmark on the internet. We need to know how the earlier data was gathered, how comparable the campaigns were, and how quickly the conditions might have changed.

We should also perform a simple sensitivity check: does the conclusion change materially under another reasonable prior? If it does, the new observations haven't overwhelmed our starting assumptions. The report should say so. Bayesian researchers emphasize that priors must be considered together with the model describing how data arise, not selected as an isolated preference.³

■ WHAT WOULD CHANGE TOMORROW?

On day three, North Pier records another batch of impressions and clicks. We don't throw away what we learned on days one and two. Under an appropriate stable-parameter model, yesterday's posterior becomes today's prior, and the new observations update it.

But we must not count the same observations twice. Nor should we assume that a campaign from six months ago and one launched yesterday are observations from an unchanged world. Sequential updating is coherent when the model and data accounting are coherent. If auction conditions, creative, attribution rules, or target audiences change, the model may need to change too.

This gives us a more demanding view of “learning in real time.” We can update whenever useful evidence arrives. We aren’t required to make a new business decision every time the dashboard changes.

That difference sounds small until there’s money on the line.

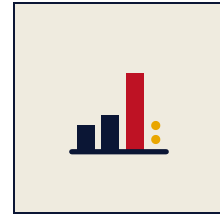
■ PRACTICE: DEFEND YOUR STARTING POINT

Before the next test, write one sentence describing what result you expect and where that expectation comes from. Give a range, not only one number. Decide what evidence would make you reconsider. Identify the business outcome that matters more than the convenient platform metric.

If an AI agent helps generate the prior or the analysis, have it show the assumptions and the calculation. Do not accept a prior merely because the agent describes it as “industry standard.” In the later chapters we’ll build a workflow in which AI can propose a model and the statistical checks can reject it.

An estimate can change with every new observation. A decision should change when the evidence and the cost of being wrong justify it.

SMALL NUMBERS, LOUD OPINIONS



At North Pier, the first two days produced three tracked orders. Had one customer purchased an hour later, the comparison would have looked different. Had one customer returned the bag, the reported advantage might have disappeared altogether.

Nothing unusual happened. Early samples are thin. The problem is our habit of treating their percentages as if they have the authority of a census.

A founder sees that one ad has converted twice as many people as another. Two orders versus one. The phrase twice as many sounds like a substantial difference. The phrase one additional person describes the same evidence with less excitement.

Both descriptions are true. Only one reminds you how easily the result could change.

■ LARGE PERCENTAGES CAN COME FROM TINY EVENTS

Suppose a landing page receives four purchases from 40 visitors. Its observed conversion rate is 10%. Another receives eight purchases from 100 visitors. Its observed rate is 8%. The first page wins the percentage contest. It does not follow that it has the higher underlying probability of conversion, or that it will win after another thousand visitors.

The samples are small, and their uncertainty is wide. A Bayesian model expresses that uncertainty as a distribution rather than forcing us to choose between 10% and 8% as fixed descriptions of the future. The point isn't that the higher number is false. The number is an accurate summary of what happened. The mistake is asking it to settle a question about what will happen next.

Here's an even more common error. A founder launches a landing page aimed at operations managers, then asks seven other founders to rate its message. The feedback may reveal awkward wording or a confusing interface. It cannot establish how operations

managers will respond as buyers. Seven people from the wrong population do not become representative because their comments are articulate.

The relevant question is not simply How much data? It's Data from whom, collected under which conditions, and about what?

One customer reports that a zipper tears during ordinary use. That's one observation. It deserves immediate investigation because it is evidence of a possible physical defect. One acquaintance says they dislike the word "durable" in the headline. That's another observation. It may be a useful prompt for a test, but not a command to change the campaign.

Counting those observations equally would be a failure of judgment.

■ SELECTION MAKES A SAMPLE LOOK BIGGER THAN IT IS

Now picture North Pier's founder reading comments on social media. The most enthusiastic customers post photographs of their bags. One angry customer writes a lengthy complaint. Everyone else remains silent.

Those comments are a sample of people who chose to comment, not a random sample of customers. The founder doesn't know whether the silent customers are satisfied, indifferent, or unhappy. If the company uses the comment section to estimate overall customer satisfaction, it has mistaken a self-selected group for the entire buyer population.

Marketing platforms create less obvious forms of selection. A platform decides who receives an impression. An algorithm may send one ad to people likely to click and the other to people likely to purchase. Later, the marketer compares the results as if the two ads had been offered to interchangeable audiences.

The totals can be correct while the comparison is misleading.

Survivorship bias appears when we study the companies that succeeded and try to identify their common habits. Perhaps the winners all used aggressive promotional tactics. We also need to know how many failed companies used exactly the same tactics. Otherwise we've described a group of survivors, not established a recipe for survival.

■ A SIMPLE EVIDENCE ASSESSMENT

Before taking action, separate four questions. What did we observe? How were those observations selected? What claim can those observations actually support? What information would most change the decision?

Applied to the friend who disliked an ad, the observation is the friend's unfavorable reaction. The selection mechanism is a personal relationship, not random recruitment. The defensible claim is that one person disliked the creative. The next useful observation would come from relevant prospective customers or from a properly designed campaign comparison.

Applied to a broken checkout, the claim is different. You have evidence of a functional defect. You can reproduce it, identify who it affects, and correct it. You don't need to estimate a population average before fixing broken software.

Small samples do not make decisions impossible. They make it essential to understand what kind of decision we're making.

Try this with a recent meeting in which someone said the data were "clear." Write the actual numerator and denominator on a page. Then write how the observations were selected. Ask whether the decision would have changed if one event had gone the other way. If it would, the conclusion may be more fragile than anyone in the room admitted.

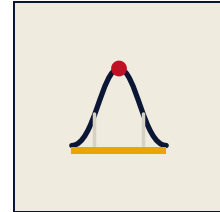


PART II

FROM STATISTICS TO A DECISION



THE WHOLE DISTRIBUTION, NOT THE AVERAGE



Return to North Pier on day fourteen. The dashboard reports 400 clicks and eight orders for Ad A; 340 clicks and eighteen orders for Ad B. Each ad spent \$500.

A clicks-to-orders conversion rate is orders divided by clicks. A's observed rate is 2%. B's is about 5.29%. Those are the recorded outcomes. To make decisions about future traffic, we have to estimate the uncertain process behind them.

For a transparent teaching calculation, use a $\text{Beta}(1,49)$ prior for each ad's click-to-order conversion rate. Its mean is 2%. We are deliberately assuming a stable conversion probability for each click and treating each reported order as an observed event. Those conditions are not always met on advertising platforms. We'll correct the measurement story later.

Under the beta-binomial model, Ad A's posterior is $\text{Beta}(9,441)$. B's is $\text{Beta}(19,371)$. The first parameter starts at 1 and receives the observed orders. The second starts at 49 and receives the clicks that didn't produce recorded orders.

A's posterior mean is 2.00%. B's is about 4.87%. B's observed rate was 5.29%, but the prior and its limited click volume draw the posterior estimate toward the starting point.

The central 95% posterior intervals are approximately 0.92%–3.48% for A and 2.96%–7.22% for B. There is uncertainty in both. Under this deliberately simple model, the posterior probability that B has the higher conversion probability is approximately 99.1%.

That's strong evidence for a difference in the rate among these observed click streams. It is not, by itself, proof that B's wording causes higher conversion, that future platform delivery will be comparable, or that B generates incremental profit.

Each of those is another question.

■ FOUR QUESTIONS THAT A REPORT SHOULD ANSWER

A posterior mean or median gives us a central estimate. The credible interval shows which parameter values remain plausible under the model. A comparison probability asks how often a draw from one posterior is better than a draw from another. Expected loss asks what we might forgo or lose by making the wrong choice.

It is the fourth question that moves the discussion from statistics into management.

Suppose two ads have a 99% probability of differing in click rate by a tiny amount. If the difference has almost no economic value, the result may not justify a new creative production cycle. Another pair of ads might show only an 80% probability of a much larger profit improvement. If switching is easy and the downside is limited, a cautious trial may be rational even at that lower probability.

The number 95 is not a natural law of marketing. It is one possible convention. A business should choose evidence requirements in relation to the stakes, not inherit them from a chart.

■ WHAT DOES A PROBABILITY OF SUPERIORITY LEAVE OUT?

Imagine two options. One has a 90% chance of earning an extra \$10 and a 10% chance of losing \$1,000. Another has a 70% chance of earning \$30 and a 30% chance of losing \$5. Choose using only “probability of being better” and you’ll miss the damage hidden in the first option.

Expected value multiplies possible outcomes by their probabilities and adds them. In this simplified example the first option has expected net value of negative \$91; the second has expected net value of positive \$19.50. A high probability of a small win is not compensation for a rare, ruinous loss.

An expected value is a useful calculation, not a complete ethics or risk policy. A startup with one payroll cycle left cannot automatically tolerate a bet merely because its average outcome is positive. Liquidity, irreversibility, and survival constraints belong in the decision.

This is the difference between a statistics report and a recommendation. The report describes plausible outcomes. The recommendation states what the company values and what risk it can bear.

■ CALCULATE WHAT CHOOSING THE WRONG AD COULD COST

We can compute the fourth question, expected loss, using North Pier rather than leaving it as a definition. Keep the Chapter Four click-to-order posteriors: A is Beta(9,441), B is Beta(19,371). Assume future cost per click stays at the day-fourteen level: \$1.25 for A and about \$1.47 for B. Each incremental order contributes \$32 before advertising. For this example only, pretend reported orders equal incremental orders and each ad will face comparable future traffic. Those are strong assumptions, and we'll challenge them later.

Suppose we must direct the next 1,000 clicks to one ad. At any candidate conversion probability p , the expected contribution for that option is $1,000 \times (\$32 \times p - \text{its CPC})$. Because p is uncertain, there is a whole posterior distribution of possible expected contributions.

Using the posterior means gives -\$610 for A and about +\$88 for B, a difference of about \$698 favoring B. That already makes the commercial meaning clearer than "99.1% chance B has the higher conversion rate." But we're not done.

For each paired draw of A's and B's conversion probabilities, ask what contribution we'd miss by committing to A when B was better, or by committing to B when A was better. Count zero regret when the chosen option is better. Averaging those amounts over the posteriors gives the posterior expected opportunity loss. The Appendix B calculation, using integration rather than a guessed percentage, produces approximately \$704.21 of expected opportunity loss if we choose A and \$5.82 if we choose B for this 1,000-click decision. B has about a 96% chance of delivering the higher expected contribution per click under these cost assumptions.⁷

Why can B still have an expected loss when it is the favored option? Because some plausible combinations of the true conversion rates leave A with the better return. A decision doesn't become risk-free merely because its expected value is highest.

The value of knowing the two conversion probabilities perfectly before this 1,000-click allocation is only about \$5.82, provided we already choose B. That's the expected value of perfect information for this narrowly defined choice. It says an expensive follow-up test would not pay for itself merely to eliminate this particular parameter uncertainty over 1,000 clicks. A larger future budget, switching penalties, or uncertainty about whether orders were actually incremental would change the calculation. Do not mistake this teaching result for permission to move real money between two nonrandomized ads.

Expected loss is not the same as the probability of loss, and neither is the same as the range of realized future cash outcomes. The calculation above compares options using their unknown expected conversion rates. Actual purchases will vary even if those rates were known. A cash-constrained company may need that predictive downside, too.

■ READ THE INTERVAL BEFORE TELLING THE STORY

There is a common mistake in executive reporting. Someone says a model estimates 3.5 ROAS, then adds that the 95% credible interval runs from 0.8 to 7.4. The second sentence nearly demolishes the certainty implied by the first. It tells us the likely range includes both serious disappointment and substantial upside.

The correct response is not to hide the interval because it makes the meeting uncomfortable. It's to name the uncertainty and identify the decision it affects.

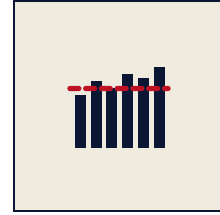
A credible interval is a probability statement about the model parameter after seeing the data, conditional on the model and prior.² It isn't a prediction that 95% of next month's observed results will land in that range. A future campaign has additional sources of variation. For that, we need a posterior predictive distribution.

You don't need to memorize the terminology to use the distinction. Ask whether the number describes an underlying rate or a possible future result. The two answers should not be quietly exchanged.

For North Pier, the immediate recommendation remains provisional. B's measured conversion rate looks better, but our comparison isn't randomized and our revenue accounting isn't yet incrementality. We should continue to investigate B's commercial advantage without representing every percentage as a cause.

Practice explaining a recent result in four sentences: what you measured, the plausible range, the chance that the relevant business threshold was met, and the cost of being wrong. If you cannot fill in the third or fourth sentence, you have found a gap in the analysis, not a communications problem.

PRIORS YOU CAN DEFEND



The phrase prior belief gives some marketers the impression that Bayesian statistics is subjective mathematics. They picture an analyst entering a preferred answer, turning a crank, and announcing a result the model was destined to find.

A badly chosen prior can do exactly that. Which is why we should make the prior inspectable.

■ A PRIOR NEEDS A SOURCE, NOT A PERSONALITY

For North Pier, we used $\text{Beta}(2,98)$ as a light CTR prior in Chapter Two. It centered the initial expectation at 2% with the influence of roughly 100 impression opportunities. It wasn't based on an actual history of insulated lunch bag ads. It was a declared teaching assumption.

In production, North Pier might have six months of ads to similar audiences under similar auction conditions. That history could justify a more informative prior. If the historical campaigns targeted different countries, used another optimization objective, or ran during an unusual promotion, we would be foolish to treat the old data as interchangeable with the new.

A prior's numerical strength must be earned by comparability.

Consider a more forceful $\text{Beta}(200,9800)$ prior. Its mean is also 2%, but it carries the concentration of 10,000 pseudo-observations under this model. Now give the ad 40 clicks in 2,000 impressions. The posterior mean remains exactly 2%. That sounds reassuring until the new ad is actually different. A very concentrated prior can hold the analysis near yesterday's answer long after tomorrow has changed.

The issue isn't that informative priors are bad. They're valuable when the underlying conditions are genuinely comparable. The problem is confusing having data with having relevant data.

■ BUILD A PRIOR FROM SIX MONTHS OF EVIDENCE

Let's give North Pier a history. This is an additional teaching dataset, not an alteration of its day-two or day-fourteen ledger. Assume six sufficiently large, comparable monthly campaign cohorts produced observed CTRs of 1.3%, 1.7%, 1.9%, 2.1%, 2.3%, and 2.7%. Assume for the moment that each month's impression count is large enough that sampling noise is small compared with the differences between months.

The mean of those six rates is 2.0%. Their sample variance, in probability units rather than percentage points, is 0.0000236. A Beta distribution has mean $\alpha/(\alpha+\beta)$ and variance $\mu(1-\mu)/(\alpha+\beta+1)$. If we match those two moments, the implied concentration is:

$$\kappa = \mu(1-\mu)/\text{variance} - 1 = 0.02 \times 0.98/0.0000236 - 1 \approx 829.51.$$

Set $\alpha = \mu\kappa$ and $\beta = (1-\mu)\kappa$. That gives approximately Beta(16.59, 812.92). This is a method-of-moments empirical-Bayes approximation: we estimated a starting distribution using historical data, rather than receiving a prior from nowhere. The exact decimals aren't sacred. The assumptions are the point.

Would you give a brand-new creative the full strength of that historical distribution? Probably not without checking comparability. Here we'll retain one quarter of its concentration as a deliberately disclosed judgment, preserving the 2% mean. That produces approximately Beta(4.15, 203.23), with concentration 207.38. Think of the discount as provisional weighting, not a law that says past data deserve exactly 25% of their original influence.

If the new ad then collects 10 clicks from 200 impressions, its raw CTR is 5%. Combining this illustrative account prior with those data gives Beta(14.15, 393.23) and a posterior mean of approximately 3.47%. The estimate moves toward the new data without pretending the first 200 impressions erased six months of potentially relevant history.

Here's the audit that matters. We have only six historical cohort estimates. Differences among their observed rates contain both real variation and sampling noise; treating all of it as variation in underlying CTR is an approximation. A serious implementation would keep denominators, model campaign-level differences, and account for uncertainty in the historical parameters. We also need to verify that the new audience and campaign objective actually resemble the old ones. The arithmetic is a way to begin eliciting a prior, not an automatic certificate that it fits.

Store this decision in a prior registry. Record the campaign population, date range, number of impressions, model and formula, historical mean and variance, discount, resulting parameters, reason for the discount, who approved it, and the date it should be reconsidered. A second analyst should be able to reconstruct the prior without guessing what "based on history" meant.

■ PRIOR PREDICTIVE CHECKS

Before seeing the current campaign results, ask what outcomes the proposed prior and data model would predict.

A Beta(1,1) prior distributes probability uniformly over click rates from 0% to 100%. In a textbook it makes the updating arithmetic easy. As a prior for an ordinary paid advertising CTR, it also puts considerable probability on results that would be astonishing. It's not an automatically sensible expression of ignorance.

A prior predictive check draws hypothetical rates from the prior, simulates impressions and clicks under the proposed data model, and asks whether those fake campaign results make sense. If the simulations routinely produce 40% click rates for a type of ad that historically draws 1%–3%, the prior or the model needs another look.

This is a good job for code, and an especially good job for an AI coding assistant supervised by someone who knows what an implausible result looks like.⁶ The assistant can simulate outcomes. It cannot decide, without context, whether your company has a reason to expect radically unusual performance.

■ TEST SENSITIVITY BEFORE ACCEPTING THE CONCLUSION

Run the same new data under two or three defensible priors: one weak, one based on genuinely similar campaigns, and perhaps one skeptical alternative. Compare the answer that matters to the business, such as the probability a conversion rate exceeds the acquisition break-even threshold.

If the recommendation changes whenever the prior changes, the current data haven't settled the issue. That is worth reporting.

If the answer stays stable across reasonable priors, we have a stronger basis for acting. It still doesn't establish that the model or measurement system is correct; it tells us one particular source of uncertainty is not driving the conclusion.

We also need to avoid double-counting. Using last month's posterior as this month's prior and then including last month's raw observations again in the likelihood counts the old data twice. It makes our estimates look more certain than the information warrants.

Similarly, pooling two years of campaign observations as though the audience and auction were unchanged is not the same thing as modeling how results have evolved. When conditions change, we may need a time-aware model or to discount older evidence. The old campaigns can inform the current analysis without pretending the world has stood still.

■ HIERARCHICAL PRIORS, IN ORDINARY LANGUAGE

Suppose North Pier launches its first ad aimed at office workers. It has little data for that particular ad but substantial history from earlier lunch bag campaigns. A hierarchical model can estimate common account-level behavior while allowing individual ads to differ. Sparse ads borrow some information from the broader account; ads with more direct evidence rely increasingly on their own observations.

This partial pooling is often preferable to starting every ad from zero or treating every ad as identical. The constructed prior above gives a small numerical version. A new ad's 10 clicks in 200 impressions suggest 5% CTR if read alone. The account-derived prior is centered at 2%. Combining the two gives about 3.47%. Full pooling would simply assign the 2% account figure to the new ad; no pooling would leave us close to its 5% observed rate. Partial pooling lies between them.

The 3.47% is a prior-updating illustration, not the result of fitting a full hierarchical model. A proper hierarchical analysis estimates how much ads vary around the account pattern rather than fixing the discount by hand. It can let extraordinary ads separate from the pack as evidence arrives. It can also hide real differences if the model forces unrelated audiences into a common pattern. Model checks should look for that possibility.

Think of partial pooling as a statistical decision about how related the campaigns are. It's not permission to transfer whatever numbers happen to be available.

A sensible prior registry records the parameter being modeled, the source data, the date, the population, the strength of the distribution, who approved it, and when it needs review. That record won't make a wrong prior right. It makes the mistake easier to discover and correct.

For the next campaign you launch, choose a starting distribution you can explain. Then name the evidence that would cause you to revise it. If your only justification is "that's the benchmark," you haven't finished the work.

A defensible prior won't rescue an experiment that was never designed to answer a worthwhile question.

BUY AN ANSWER WORTH KNOWING



A client asks a marketer to test two headlines. The marketer says yes, creates the ads, and spends \$300. A week later the client wants to know which won.

The marketer has no persuasive answer. The test produced a handful of orders. It was never designed around a difference small enough to matter, or large enough to detect with the proposed budget. Nobody settled what “won” would mean before the money went out.

That’s not an unlucky experiment. That’s an experiment that wasn’t planned.

■ BEGIN WITH THE DECISION, NOT THE SPLIT

For North Pier, the real choice might be whether to move the next \$5,000 of advertising spend toward message B. The business question is whether B produces more incremental contribution after the added advertising cost, not simply a higher CTR.

Before running a controlled test, state the outcome that matters, the smallest improvement worth paying attention to, the eligible population, how assignment will occur, and the point at which we will review evidence. This doesn’t require pages of bureaucracy. It requires making commitments before results tempt us to change the rules.

Suppose we’re evaluating click-to-order CVR and would only make a substantial change if conversion improved from 3% to 4%. That’s a one-percentage-point absolute lift, or roughly 33% relative lift. Those are two ways of describing the same proposed change, and it’s easy to confuse them.

For orientation, a conventional two-sided fixed-sample comparison with 5% type-I error and 80% power would require roughly 5,300 clicks per arm to distinguish rates of 3% and 4%, under standard assumptions of independent Bernoulli outcomes and balanced

assignment. It's an approximation, not a universal Bayesian minimum or a promise that a particular test will be conclusive. The required sample changes with the underlying rate, the business threshold, the test method, dependence among observations, and the required precision.

North Pier's day-fourteen sample has 400 clicks on A and 340 on B. If management had aimed to detect a change that small, it would have been unrealistic to budget the experiment as though two weeks of similar traffic were enough.

■ WHAT DOES KNOWLEDGE COST?

Suppose a click costs \$1.40 and we require about 10,600 total clicks across both versions. That's approximately \$14,840 in traffic cost, ignoring production costs and auction changes. For a startup selling one product, that may be an irrational price to pay for detecting a relatively small CVR difference. It might be cheaper to test a larger message change, find a more frequent intermediate outcome that validly predicts future purchases, or decide that the proposed change isn't worth resolving at all.

Don't confuse a cheaper proxy with a substitute for actual sales. If a landing page attracts more clicks but those clicks don't turn into incremental contribution, you've optimized the wrong target. A useful proxy must have a demonstrated connection to the outcome you care about.

Now we can ask the question most test briefs omit: What is this answer worth?

If a better allocation is expected to improve future contribution by \$50,000, spending several thousand dollars to distinguish the alternatives may be sensible. If the entire decision could move contribution by only \$1,500, purchasing a \$14,840 answer would be absurd.

A formal expected value of information calculation compares the best expected result obtainable now with the best expected result after collecting additional evidence, averaging over the possible evidence we might receive. We subtract the cost of obtaining it. The value depends on the choices, their payoffs, and uncertainty.⁷ It is not an intrinsic property of a sample size.

A smaller experiment may still be worthwhile because it can eliminate obvious problems or improve the next design. But we should be explicit: learning something is not the same as learning enough to justify the intended decision.

■ WORK OUT THE VALUE OF A TEST BEFORE BUYING IT

A compact example makes the calculation tangible. This is a separate hypothetical choice, not North Pier's measured campaign. We may keep the current allocation, with an additional payoff of \$0, or switch to a new approach. If the approach really works, switching yields \$2,000 in future contribution compared with holding. If it doesn't, switching costs \$1,000. Based on what we know now, assign a 40% chance that it works and a 60% chance it doesn't.

Without more evidence, the expected benefit from switching is $0.40 \times \$2,000 - 0.60 \times \$1,000 = \$200$. So switching is the better available decision on expected contribution. If we somehow knew the true state perfectly, we'd switch in the successful 40% of cases and hold in the others: $0.40 \times \$2,000 = \$800$. Perfect information would be worth $\$800 - \$200 = \$600$. That is an upper bound on the value of an imperfect test for these two choices and payoffs.

Now suppose a proposed test has a 75% chance of returning a positive result when the new approach works, and a 25% chance of returning a positive result when it doesn't. These are explicit assumptions about the test, not magical properties of an A/B test. The probability of seeing a positive result is $0.40 \times 0.75 + 0.60 \times 0.25 = 45\%$. After a positive result, the probability the approach works becomes $0.30/0.45 = \text{two thirds}$. At that point switching has an expected benefit of \$1,000. A negative result lowers the probability of success to $0.10/0.55 \approx 18.2\%$, making holding preferable.

Follow that decision rule across both possible test results. The expected contribution after receiving the test result but before paying for the test is $0.45 \times \$1,000 + 0.55 \times \$0 = \$450$. Compared with the \$200 available without testing, the expected value of sample information is \$250. If the test costs \$150, its expected net value is \$100. If it costs \$350, don't purchase this test on these assumptions.

What changed? Not the statistical vocabulary. We priced the uncertainty against a real choice and the amount at stake. We also had to justify how informative the proposed experiment would be. An AI-generated test plan that cannot support those assumptions has no business printing "EVSI = \$250" as a factual prediction.⁷

■ RANDOM ASSIGNMENT IS DOING WORK THE MODEL CANNOT DO

North Pier's original comparison consisted of two ads separately delivered by a platform. If different types of users received the messages, the observed difference may be partly due to the selection of recipients rather than the message itself.

In a genuine experiment, eligible people or other appropriate units are assigned to conditions by a controlled random process. The assignment tries to make the groups comparable before they see the treatment. In advertising, implementation is harder than it sounds. We may need platform-native experiments, mutually exclusive audience splits, or geo-level assignment. We must avoid users appearing in both groups and take account of campaign delivery algorithms, budgets, and conversion delays.

Randomization doesn't guarantee a perfect experiment. It does make a causal comparison more credible than comparing two independently optimized ad sets and assuming the only difference is the text.⁵

■ LOOKING AT RESULTS EVERY MORNING

Many marketers were taught never to inspect a test before a fixed endpoint. That's a poor simplification. We can monitor safety, data quality, delivery, and operational failures from day one. Bayesian updating also permits coherent sequential inference under an appropriate model. Frequentist approaches have valid group-sequential and always-valid methods when specified and implemented properly.

The problem is not seeing data. It is repeatedly using the same conventional fixed-sample stopping threshold as if we had made only one test at one predeclared moment.

Conversion delay creates another trap. Someone clicks on Thursday but buys on Saturday. A Friday report shows the click without the eventual order. Treat it as a completed failure and you're manufacturing apparent underperformance from incomplete observation.

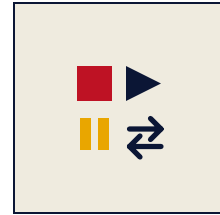
Google's own Smart Bidding guidance cautions against evaluating changes without accounting for learning and conversion delays.⁸ That's a platform-specific operating concern, not proof that the platform needs to be left untouched under all circumstances.

Before spending, give the test a decision rule and a calendar for review. Define exceptions for defects, severe financial harm, or other early-stop conditions. Then let the evidence accumulate according to that rule rather than changing it whenever the dashboard

becomes uncomfortable.

A test is not a ceremony. Its purpose is to improve a decision enough to justify its cost.

STOP, CONTINUE, HOLD, OR SWITCH



A Bayesian report says the probability B has a higher conversion rate is 99.1%. An executive wants to move the entire budget immediately.

Should we?

We still need to know whether the data reflect comparable audiences, whether the improvement will persist at a higher spend level, what switching will cost, and how much error the company can bear. The 99.1% number answers one particular rate comparison under a particular model. It doesn't sign the budget approval.

■ DECISIONS HAVE ASYMMETRIC COSTS

Consider a separate simplified example with two possible choices. If we keep the current campaign, expected contribution for the next week is \$1,000. If we switch, there is a 70% chance of earning \$1,500 and a 30% chance of earning \$400. The switch has expected contribution of \$1,170 before switching costs. If switching costs \$250, the expected result falls to \$920. Holding becomes the better choice under these assumptions, even though switching often produces a higher outcome.

This illustration doesn't require Bayes to do the arithmetic. Bayesian modeling becomes useful when it supplies an updated probability distribution for the outcomes rather than pretending that the 70% arrived from nowhere.

Budget decisions should compare posterior expected values after costs, and sometimes consider a credible downside or a survival constraint. The correct action may be to expand B cautiously rather than reassign everything.

■ NORTH PIER'S BREAK-EVEN QUESTION

Under the original case assumptions, each incremental order contributes \$32 before advertising. At day fourteen, A's observed cost per click is \$1.25. B's is about \$1.47. If those CPCs were stable, an ad would need an incremental click-to-order conversion probability exceeding CPC divided by \$32 to cover the next click's cost.

For A, that threshold is approximately 3.91%. For B it is approximately 4.60%.

Under the simplified click-to-order posteriors from Chapter Four, A has less than a 1% posterior probability of exceeding its threshold; B is around 57%. These are calculated probabilities for expected contribution per incremental click, conditional on fixed future CPC, equal \$32 contribution per incremental order, and directly observed, causally incremental conversions. They are not probabilities of next month's profit. Real auctions change with spend, and platform attribution is not incrementality.

Notice what this does to the earlier celebration of B. B is a far stronger candidate than A on the observed funnel. Yet under our assumptions, the posterior probability that B's expected per-click contribution clears break-even is only a little better than even.

We are not finished learning.

■ HOLDING CAN BE AN ACTIVE DECISION

A pause in intervention is not the same as a pause in observation. We can leave the allocation steady, keep collecting data, monitor anomalies, and improve the test design while avoiding expensive alterations triggered by noise.

This matters with automated platforms. Frequent changes to objectives, budgets, or targeting can alter delivery conditions, complicating comparisons before and after the change. No universal formula tells us the exact cost of touching a campaign. We need account-specific evidence and the platform's current rules.

There are cases when holding is reckless. If an ad makes a false claim, correct it. If conversion tracking has failed, repair measurement before trusting the report. If the business can't survive another week of losses, do not wait for statistical elegance while the cash disappears.

A rational decision rule considers the consequences of delay as well as the consequences of a mistaken action.

■ A DECISION RECORD THAT OUTLIVES THE MEETING

Before changing North Pier's allocation, document the existing evidence, the action under consideration, the expected benefit, the downside, the conditions that would reverse the choice, and the date of review.

For example: "We will run a controlled follow-up with B as the favored candidate. We will not interpret day-fourteen observational differences as causal proof. The next test will compare incremental orders or contribution, preserve a simultaneous control, and account for conversion delay. If the experiment's projected value of information exceeds its cost, we will fund it; otherwise we will choose a bounded allocation and keep monitoring."

Not a slogan. A decision someone can examine after the outcome arrives.

This is also the best answer to a founder who insists every meeting needs an action item. The action may be to maintain the present setup until a predeclared condition is met. It may be to repair the measurement. It may be to stop the test because the result isn't commercially worth resolving.

Bayesian statistics helps estimate the uncertain outcomes. It doesn't absolve us from deciding what we are willing to lose.

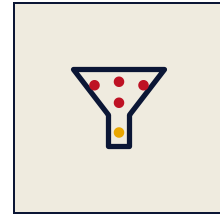


PART III

THE BUSINESS BEHIND THE DASHBOARD



ONE FUNNEL, SEVERAL LINKED QUESTIONS



The ordinary marketing dashboard divides a business into columns. Impressions. Clicks. Conversions. Cost. Revenue. Return on ad spend.

The columns are useful. They are also deceptive when we treat each one as a separate contest.

An impression can become a click. A click can become an order. An order has a value and a cost. A campaign pays for delivery throughout. The metrics describe related events in one process.

If you improve CTR by attracting people who aren't likely buyers, you may increase traffic and lower commercial performance at the same time.

North Pier has already experienced precisely that pattern.

■ FOLLOW THE NUMBERS FROM TOP TO BOTTOM

The day-fourteen ledger is straightforward:

MEASURE	AD A	AD B
Impressions	20,000	20,000
Clicks	400	340
Tracked orders	8	18
Spend	\$500	\$500
Tracked revenue	\$640	\$1,440
Observed CTR	2.00%	1.70%
Observed click-to-order CVR	2.00%	5.29%
Spend per tracked order	\$62.50	\$27.78
Tracked revenue ROAS	1.28	2.88
Contribution after spend, using tracked orders	-\$244	\$76

The last row uses the \$32 assumed first-order contribution per tracked order before ad spend. It isn't verified incremental profit. We keep repeating that limitation because it changes which business decision the numbers can support.

A's higher CTR does not rescue its poor tracked contribution. B's stronger lower-funnel result does not establish that its message caused the advantage. Both statements can be true.

■ WHY FOUR INDEPENDENT MODELS ARE NOT ENOUGH

A simple teaching model might place a Beta distribution around CTR and another around CVR, then estimate CPA and ROAS from the resulting point estimates. That can help explain updating, but it can miss dependence. A creative can attract a particular kind of clicker; the distribution of clickers affects subsequent conversion. Order values vary. Spend responds to auctions and targeting choices.

If we freeze revenue per order at a single number when actual order value is variable, the model makes ROAS seem more certain than it is. If we treat clicks and conversions as unrelated independent outcomes, we can lose information about how traffic quality and

purchase behavior interact.

A joint generative model describes how observations might be produced together. It can share parameters or correlated effects across funnel stages and derive CTR, CVR, CPA, and ROAS from simulated possible funnel states. This allows uncertainty in conversion rates, purchase values, and relevant cost components to move through to the quantities decision-makers actually use.

The word generative means the model can produce hypothetical datasets under its assumptions. That ability is useful for checking whether it can reproduce important features of the observations, not just return attractive averages.

We should be precise about “joint.” Merely placing independent models in the same report does not make them coherent. Conversely, a joint model does not become correct by being elaborate. Wrong assumptions can create synchronized errors across every reported metric.

■ A JOINT MODEL CAN CHANGE THE ANSWER

Try a small simulation you can check without software. This is a hypothetical account, not North Pier’s observed data. Suppose its ads can run in two equally likely audience conditions. In a curiosity-heavy condition, CTR is 4% but only 1% of clicks purchase. In a purchase-ready condition, CTR is 1% but 5% of clicks purchase. Each condition generates 10,000 impressions; spend is held at \$200 and an order brings in \$80.

In the curiosity condition, we expect 4 orders, \$320 in revenue, and 1.6 ROAS. In the purchase-ready condition, we expect 5 orders, \$400, and 2.0 ROAS. Given equal probability, the expected ROAS is 1.8. That calculation preserves the relationship between who clicks and who buys.

Now commit a common modeling mistake. Take the marginal CTR values (4% or 1%) and the marginal CVR values (1% or 5%) and randomly pair them as though the two stages were independent. You invent four equally likely combinations, including the fantasy of both 4% CTR and 5% CVR. Those four ROAS outcomes are 0.4, 1.6, 2.0, and 8.0. Their average is 3.0. The original paired scenarios, by contrast, allow only 1.6 or 2.0.

The independent construction says average ROAS exceeds North Pier’s illustrative break-even threshold of 2.5, and gives that imagined 8.0 outcome a 25% chance. The actual paired conditions never clear 2.5. That’s a materially different decision from the

same marginal summaries.

This is an example of negative dependence. It doesn't prove CTR and CVR always move in opposite directions. Sometimes they move together, sometimes the dependence is weak, and observed traffic mix is often determined partly by the delivery system. A joint model must learn an appropriate relationship from relevant evidence; simply declaring two stages correlated would be another unsupported assumption. For a quick single-rate sanity check, separate Beta models can still be useful. They become unsafe when we multiply their independently sampled outputs into a profit claim without checking the dependence and the uncertainty in orders, value, and spend.

■ PARTIAL POOLING WITHOUT PRETENDING ALL CUSTOMERS ARE ALIKE

An agency may have dozens of campaigns for related products. A brand-new ad has only ten clicks; another has 10,000. Should the new ad's early rate be treated as if it were wholly unrelated to the account's history? Probably not. Should the new ad be assigned the account average as if it were known to be identical? Also no.

A hierarchical model lets campaign- and account-level patterns inform sparse ads while allowing the ads to differ. This is partial pooling. The amount of sharing is itself informed by the data and the model.

But imagine pooling North Pier's urban office campaign with a summer beach campaign, using the same assumptions despite obvious differences in intent and season. The model may produce precise estimates by erasing genuine distinctions. That is over-pooling.

Good analysis checks whether the pooled model can reproduce the variability between ads and campaigns. It also compares sensible alternatives: fully pooled, partially pooled, and separate estimates.

■ REAL DIAGNOSTICS, NOT COMFORTING CHARTS

A Bayesian computation should report more than parameter estimates. It should show whether the calculation itself was trustworthy.

For a model fitted by Markov chain Monte Carlo, diagnostics such as split-chain R-hat, effective sample size, and divergent transitions help detect computation problems.⁹ R-hat compares how much several chains disagree with one another against their varia-

tion inside each chain. Near 1.00 is desirable; below about 1.01 is a commonly used working criterion, not a law. Effective sample size, or ESS, estimates how many independent draws the correlated simulated draws are worth. Hundreds of effective draws may be a practical starting check for important parameters, but tail estimates, precision demands, and the model matter. A divergence is the sampler reporting that its numerical trajectory could not reliably explore part of the posterior geometry. Don't wave it away because the plot looks pretty. Investigate and refit the model.¹⁷

Posterior predictive checks answer a different question: can simulated data from the fitted model reproduce important features of the actual observations? Cross-validation asks how well the model predicts held-out observations or how candidate models compare. None of these checks proves a causal mechanism. A model can converge perfectly to the wrong answer if the likelihood is wrong or the data are misleading.

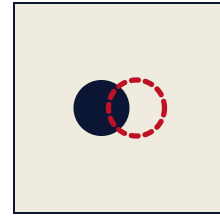
Those checks answer different questions. A sampler that converges has not proved the business model correct. A model that predicts a held-out sample well has not necessarily isolated the causal effect of advertising. And a chart of independently drawn numbers from a closed-form distribution is not a convergence test for a sampler that never ran.

This distinction is central to the newer Bayesian Marketing Decision Framework. Its proposed rebuild replaces isolated per-metric updates with a joint funnel model, explicit data contracts, genuine diagnostics, and derived business quantities. Those are design goals described in its architecture and specification, not verified capabilities we should pretend are already deployed.

For the working marketer, the practical question is simpler: Can the analyst show how the number was derived, what assumptions made it possible, and what checks would reveal a failure?

If not, the output may be polished. It isn't ready to govern a budget.

REPORTED IS NOT CAUSED



North Pier's dashboard shows eighteen orders attributed to Ad B. The founder wants to celebrate eighteen new customers.

Were they new because of the ad?

Some may have purchased anyway. Others may have seen the advertisement, searched for the brand later, and purchased through a different channel. A few purchases may have been attributed to multiple campaigns. Still others might be missing from the platform report altogether.

The ad platform reports what its measurement system associates with the campaign. Incrementality asks how many additional orders occurred because advertising ran, compared with what would have happened without it.

Those are not the same number.

■ THE COUNTERFACTUAL WE CANNOT DIRECTLY OBSERVE

Take one customer. They bought a North Pier bag after seeing an ad. Did the ad cause the purchase? To know for certain, we'd need to observe that exact customer in an otherwise identical world where the ad never appeared. We cannot observe both realities for the same person at the same moment.

A controlled experiment gets around that impossibility by comparing groups assigned to advertising and no-advertising conditions.¹⁰ Random assignment helps make the groups comparable before treatment. The difference in outcomes can then estimate average causal impact for the tested population, subject to compliance, interference, measurement quality, and other assumptions.

A holdout group receives no campaign exposure under the experiment's specified conditions. A geo test assigns comparable geographic areas to treatment and control. A conversion-lift study aims to estimate additional conversions attributable to the campaign.¹³ Each method has limits. Poorly chosen groups, spillover between groups, weak power, and missing purchases can still damage the estimate.

Marketing mix modeling looks at patterns across time, channels, and business conditions. Modern Bayesian MMM tools can incorporate prior information, model carryover effects, and use experimental results to calibrate estimates. They do not make confounding vanish merely because they are Bayesian.

Google's Meridian documentation, for example, describes using incrementality experiments to calibrate return-on-investment priors.¹² The calibration is an attempt to bring experimental evidence into the broader model. It is not a claim that aggregate modeling is equivalent to random assignment.

■ WHY 2.88 ROAS MAY STILL BE THE WRONG ANSWER

North Pier's reported B ROAS is 2.88. For each \$1 of advertising spend, the platform attributes \$2.88 in revenue. Under the simplified first-order margin of 40%, \$2.88 in revenue generates about \$1.15 of pre-ad contribution. Subtract \$1 of ad spend and the remaining contribution is around \$0.15.

That's why the case's break-even ROAS is 2.5, not 1.0. A revenue ROAS of 1.0 returns \$1 of sales against \$1 of ads while still leaving production, fulfillment, processing, and returns costs unpaid.

But even 2.88 is not proof of incremental profit. Suppose, in a clearly hypothetical alternative scenario, a rigorous lift test estimates that only twelve of the eighteen tracked orders were additional sales caused by the campaign. Then the test's incremental first-order contribution is $12 \times \$32 = \$384$ against \$500 in spend. On that estimated basis, the campaign loses \$116 before fixed overhead.

We aren't rewriting North Pier's historical case; we're illustrating how an incrementality estimate could change its economic interpretation.

There may be customer lifetime value beyond the first order. There may also be refunds, discounts, support expenses, and acquisition of existing customers who would have returned anyway. The right financial model depends on what the business actually

earns, not on an impressive return ratio.

■ WHAT THE PLATFORM KNOWS, AND WHAT IT GUESSES

Ad platforms may report modeled conversions, aggregated events, conversions assigned through attribution rules, or delayed events that later revise an earlier report. Their numbers can be useful operational signals. They aren't always independent direct observations of purchases.

If a dataset contains a fractional reported conversion count, such as 12.7, we should not quietly feed that number into a binomial likelihood meant for a literal count of independent yes-or-no purchase events. We need a measurement model appropriate to how the fractional number was produced, or a different data source for customer-level outcomes.

Attribution windows also matter. A seven-day click window and a one-day click window may associate different sales with the same campaign. If the window changes between reports, the apparent performance change can reflect the measurement rule rather than customer behavior.

A useful report states which conversions were observed, which were modeled, what attribution rules were used, and what remains outside measurement.

You can do excellent Bayesian inference on the data that survive those rules. You cannot recover information that was never observed without additional assumptions or evidence.

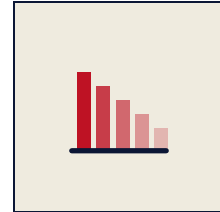
■ A PRACTICAL MEASUREMENT LADDER

Start with clean transactional data if available. Reconcile orders, refunds, dates, currencies, and duplicate identifiers. Understand the platform's reported conversions and why they differ from the business ledger. Use platform metrics to guide operational questions while reserving causal language for designs that support it.

When the decision justifies the cost, run a randomized holdout or a sound geo experiment. Use its results to calibrate broader models rather than treating a dashboard attribution count as an experiment. For a small startup, a full MMM may be premature; simple randomized tests and honest accounting may produce more useful answers for less effort.

The point isn't to become suspicious of every metric. It's to know which question each metric actually answers.

WHEN YESTERDAY STOPS DESCRIBING TODAY



North Pier learns from fourteen days of advertising. The following month, its click costs rise. A holiday promotion begins. The platform introduces a delivery change. A competitor offers an almost identical bag at half the price.

Should last month's results carry the same weight in the next decision?

Not automatically.

Bayesian updating is coherent when the model correctly describes how observations are generated. If the environment changes, treating old and new data as independent draws from one unchanged process may manufacture certainty.

■ THE CAMPAIGN HAS A CLOCK

Creative fatigue changes how people respond to repeated impressions. Seasonality changes demand. Promotions alter purchase intent and order value. Auction competition changes which impressions the budget can buy. Tracking systems change attribution and the dates on which conversions become visible.

Different forces have different timescales. A minor daily fluctuation may be noise. A permanent shift in the purchase process may require a new model. Not every variation deserves a pivot; not every persistent deterioration can be explained away as a slow week.

One approach is to include time in the model: allow conversion probability, click cost, or revenue per order to vary over days and weeks. A state-space model estimates a changing underlying state rather than forcing one fixed parameter to describe the whole

period. Another approach is to model known drivers such as weekdays, promotions, placements, or season. A simpler analysis may divide clearly different periods and avoid pooling them as though nothing changed.

Each choice trades flexibility against the risk of overfitting. A model that can explain every daily wiggle after the fact may be poor at predicting tomorrow.

■ DECAY CAN BE RATIONAL, BUT IT HAS TO BE MODELED

Suppose North Pier has 200,000 historical impressions from a summer campaign and 2,000 new impressions from an autumn campaign. If it mechanically combines both periods under one stable CTR assumption, the old data dominate even when the product, audience, and competitors have changed.

A time-aware model can reduce the influence of older observations or allow parameters to evolve. But “let’s give last month’s results half the weight” is not automatically a principled solution. It is a modeling choice that needs justification and sensitivity testing.

Historical evidence should inform a new campaign in proportion to its relevance. This is the same principle we used for priors, now complicated by time.

■ WHAT CHANGED: BEHAVIOR, DELIVERY, OR MEASUREMENT?

When a result moves sharply, investigate three possibilities before deciding the product has stopped working.

First, customer behavior may really have changed. Perhaps the market is saturated, competitors improved, or the weather made the insulated bag less relevant.

Second, delivery may have changed. Auction prices, placements, frequency, geographic mix, and platform optimization can alter the customers who encounter the creative.

Third, measurement may have changed. An API field, attribution window, consent rate, conversion definition, or reporting delay can create a discontinuity that looks like behavior.

The distinction is practical. Fixing measurement won’t restore declining customer demand. Changing creative won’t correct a broken conversion event. Increasing budgets won’t compensate for the disappearance of a high-intent audience.

A versioned data contract helps. Record what each field means, when a platform definition changed, which time zone the report uses, whether a conversion may arrive late, and whether a transaction can be counted twice.

The marketing team needs a memory that is more reliable than the dashboard's current view.

■ WALK THROUGH A PERFORMANCE DROP

Suppose a separate illustrative campaign, not North Pier's ledger, records 1,000 clicks and 40 platform-attributed orders one week. The next week it records 1,000 clicks and 20 orders. Its dashboard CVR has apparently fallen from 4% to 2%. The request arrives immediately: replace the creative.

Start with measurement. Did an attribution window shorten from seven days after a click to one day? Did the conversion pixel stop firing on a payment path? Compare the underlying order system, conversion timestamps, attribution definitions, and reports rebuilt using a consistent window. If the second week's actual orders are comparable but late conversions are missing, the dashboard drop is not evidence that customers stopped buying.

If measurement holds up, inspect delivery. Here's a numerically separate example of how audience mix can create a decline with no change in behavior inside either group. During the first week, 800 clicks went to a high-intent placement converting at 5%, and 200 went to a broad placement converting at 1%. That's 42 orders from 1,000 clicks, or 4.2%. In the second week, only 300 clicks reached the high-intent placement and 700 reached the broad one. At the same within-placement rates, the total is 22 orders, or 2.2%. The apparent collapse follows directly from the delivery mix. It doesn't yet show that either group became less willing to buy.

If the measurement channel and relevant delivery composition are stable, investigate behavior: actual conversion changes within comparable cohorts, competitive offers, pricing, page defects, and creative fatigue. In real campaigns these causes can coexist. The sequence isn't an automatic diagnosis. It's a guard against telling a behavioral story before checking what was actually measured and to whom the platform delivered it.

Historical evidence can also lose relevance. One transparent starting sensitivity analysis assigns older observations a weight that halves every 30 days: $\text{weight} = 2^{-(\text{age in days} / 30)}$. A 30-day-old observation gets weight 0.5; a 60-day-old one gets 0.25. This is a proposed decay model, not a fact about ad auctions. Compare it with no decay and with half-lives such as 14 or 60 days using rolling historical forecasts. Choose only if it improves prediction and calibration on later periods. An actual time-varying model is often preferable. You cannot recover missing conversions, cure confounding, or obtain valid causal lift merely by discounting old rows.

■ FORECASTS SHOULD ADMIT WHEN THE WORLD HAS MOVED

A good forecast produces a range, not a guarantee. After the forecast period ends, compare what actually happened with what the model predicted. Did the outcomes fall within expected ranges at the expected frequency? Do particular seasons or audiences repeatedly break the model? Are intervals systematically too narrow?

This is where posterior predictive checks and out-of-sample assessment matter. If the model repeatedly fails under new conditions, it must be revised rather than defended because it once fitted historical data beautifully.

The question isn't whether to trust yesterday or today. It's whether the assumptions connecting them still hold.

Try reviewing one old campaign forecast. Don't begin by asking whether the prediction was "right." Ask what changed between the period used to fit the model and the period it attempted to predict. You may discover that the problem was not the mathematics but an assumption nobody remembered making.

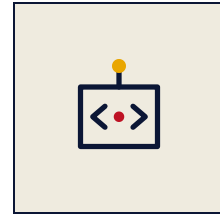
IV

PART IV

AI AND PROFESSIONAL JUDGMENT



AN AI ANALYST WITH A SUPERVISOR



A marketer who could not write a statistical model five years ago can now ask an AI coding agent to produce one. The agent can generate Python, prepare a table, select a probability distribution, sample a posterior, and describe the output in reassuring prose.

It can also get almost every step wrong.

The first result may still look professional.

This makes AI genuinely useful and potentially dangerous. The barrier to producing an analysis has fallen faster than the barrier to recognizing a bad analysis.

■ SEPARATE THE QUESTION FROM THE CODE

Let's give AI a specific North Pier task: "Estimate the probability that the true click-to-order conversion rate for Ad B exceeds Ad A under a stated beta-binomial model. Use the day-fourteen clicks and orders, show the prior, compute uncertainty, and do not describe the result as causal lift."

That's a legitimate job. The data are supplied, the estimand is defined, and the assumptions are bounded. A coding agent can produce code using a numerical library, show how it updates Beta distributions, draw samples with a fixed random seed, and calculate the comparison probability. We can test the result against a known analytic formula or independent implementation.

A bad prompt would be: "Tell me which campaign will make more money next month." The agent might respond with a confident answer without identifying the missing information about future costs, attribution, customer value, and delivery.

The improvement isn't a more eloquent prompt. It is asking a question the available evidence can answer.

■ THE SUPERVISED WORKFLOW

I use a simple principle in my own system architecture: AI proposes; verification gates decide. That design can be applied even by a small team without custom infrastructure.

First, give the assistant an explicit data dictionary and a narrow question. Have it inspect rows, identify missing values, reconcile totals, and flag impossible combinations such as clicks exceeding impressions. Keep the original data unchanged and log every transformation.

Second, request a model specification before code. Which variable is the outcome? Which population is being estimated? What assumptions connect the observed data to the unknown quantity? What prior is proposed and why? How would an alternative model change the conclusion?

Third, execute the statistical calculation in reproducible software. Do not ask a text model to invent a posterior interval. Let the numerical library calculate it. Fix random seeds for checks when stochastic computation is involved, and record package versions.

Fourth, inspect both numerical and conceptual diagnostics. For a simple conjugate beta-binomial update, check the algebra directly. For a complex hierarchical model, inspect chain convergence, effective sample sizes, divergent transitions, prior predictive simulations, and posterior predictive fit. If the model doesn't reproduce important properties of the observed data, good convergence is not enough.

Fifth, give the AI the verified outputs and ask for a plain-language explanation. Check every quantitative claim against the calculation. Don't allow the narration to promote a model-based estimate into causal certainty or turn "insufficient evidence" into a sales-ready recommendation.

Sixth, put a human name on the decision. The assistant didn't choose the risk tolerance or promise the company could survive a failed allocation. Those responsibilities belong to someone accountable.

■ TOOLS WORTH KNOWING IN 2026

PyMC provides probabilistic modeling in Python. PyMC-Marketing builds on that ecosystem with models for marketing mix, customer lifetime value, customer choice, adoption, and media budgeting.¹⁴ The useful distinction is not that the package is auto-

matically right; it is that it supplies open, inspectable implementations and workflows that an analyst can test and adapt.

Google's Meridian is an open-source Bayesian marketing mix modeling framework.¹¹ Its documentation covers model specifications, calibration using experimental results, diagnostics, and media outcome estimation. It addresses a different scale of question from comparing two ads' click-to-order conversion rates. You need enough credible data and a sensible causal design for an MMM to answer useful business questions.

Meta's Robyn offers open-source, semi-automated marketing mix modeling using methods including regularized regression and automated optimization.¹⁵ It isn't simply another name for a Bayesian model. Methodological distinctions matter, especially when comparing assumptions and uncertainty reporting.

A general coding agent can work alongside these systems, read documentation, generate scripts, and automate report assembly. But the agent must not be treated as an independent source of truth about what the packages calculated.

Current tool names, versions, APIs, and documentation can change.¹⁶ Verify the installation and the statistical behavior on the day you publish or deploy a workflow.

■ COPYABLE INSTRUCTION FOR AN AI CODING ASSISTANT

The following is a reusable analysis request, not a substitute for reviewing the result:

PROMPT

You are assisting with a marketing experiment, not making the business decision. Use only the attached data and explicitly identified prior information. Begin by defining the estimand, unit of analysis, observation window, and assumptions. Validate denominators, duplicates, missing values, currency, and conversion definitions. Propose one simple statistical model and one credible alternative. Produce executable code for the calculations, tests against a small known example, numerical diagnostics appropriate to the model, and an output table. Do not claim causal lift without a valid design. Report the probability of relevant business outcomes and the sensitivity to assumptions. If data are insufficient for the requested decision, identify the missing evidence rather than creating an answer.

Then ask a human analyst to check the code and outputs. If no analyst is available, keep the decision narrow, the model simple, and the stakes small enough that a mistake is survivable.

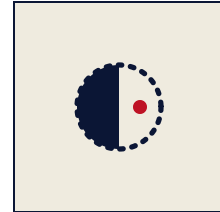
■ WHAT AI CANNOT RECOVER

An assistant can help detect a missing attribution field. It cannot reconstruct purchases that were never recorded without introducing assumptions. It can write a sophisticated model of spend, impressions, and conversions. It cannot turn nonrandom platform delivery into a randomized experiment by giving the model more parameters.

It can explain a theorem. It cannot demonstrate that a theorem about a theoretical allocation problem guarantees your live advertising revenue without checking every assumption that connects them.

Sometimes the correct product behavior is to refuse a requested probability because the input doesn't support it. That's a feature, not a failure of ambition.

BUILD A PRACTICE THAT CAN SAY “WE DON’T KNOW”



North Pier has reached the end of its first meaningful review cycle. Ad A won on CTR. Ad B produced more tracked orders and better first-order contribution under the case assumptions. A simple beta-binomial comparison provides strong posterior evidence that B’s observed click stream has a higher conversion probability.

The founder wants a decision.

There is enough evidence to stop treating A’s higher click rate as proof of better marketing. There is not enough evidence to say B’s words caused its superior commercial result, nor that the observed attributed contribution will survive more spending.

What would a serious marketing team do?

■ THE MEETING SHOULD HAVE ONE PAGE WORTH DEFENDING

I’d put the business question at the top: “Should North Pier favor B for the next allocation, and what further evidence would justify scaling?”

Then the accounting: B’s tracked first-order contribution after \$500 in spend is \$76 under the \$32-per-order assumption. A’s is negative \$244. Both use attributed orders and exclude many real-world cost and attribution complications.

Next, the statistical distinction: our simple posterior puts B ahead on click-to-order conversion for the observed streams with about 99% probability. Its probability of exceeding a simplified expected per-click break-even threshold, under fixed CPC and fully incremental orders, is only about 57%. That gap is a reminder that relative superiority and absolute business viability are different questions.

Then measurement: the delivery wasn't randomized, and tracked orders have not been shown to be incremental. A platform change or higher auction cost could alter the result. We should check conversion lag, reconcile customer orders, and validate how each creative was delivered.

Finally, the decision: favor a bounded, controlled follow-up of B rather than treating CTR as the deciding metric or moving the entire budget on observational evidence. Establish how much information the next test is expected to purchase, and define the financial downside North Pier can tolerate before changing the allocation.

The founder can disagree with that recommendation. But they can now disagree with a visible chain of evidence rather than with an analyst's mysterious number.

■ MAKING UNCERTAINTY ORDINARY

Most campaign meetings are arranged to reward certainty. The person with the strongest-sounding recommendation gets the room. The analyst who reports a wide interval is asked to "tighten it up" as if uncertainty were a formatting choice.

That is how bad reporting becomes a habit.

A better practice records what we thought would happen before the experiment, the conditions under which we expected it, and what we would do if the evidence went the other way. When a test ends, the team compares the result with the expectation, not with a story written after everyone knows the outcome.

It also keeps the questions separate. What happened? What do we infer about the uncertain underlying process? What caused what? What action has the best expected result given the company's priorities and constraints?

One meeting can answer some of those questions and leave others open. That's allowed.

■ SEVEN QUESTIONS BEFORE YOU APPROVE A RECOMMENDATION

Where did this number come from? Which model produced it? Did the computation pass its checks? What does the data actually permit us to claim? What would acting cost if we're wrong? Who takes responsibility for the decision? Can we return later and compare the outcome with what we believed today?

Those are the same questions whether the report comes from a spreadsheet, a hired analyst, an AI agent, or a future decision system. They're worth keeping in the final chapter because they belong to the professional making the call, not to any particular product. A description of one possible software architecture appears in Appendix G for readers interested in how to enforce some of those checks with code.

■ A DIFFERENT KIND OF EXPERTISE

Marketing will always involve incomplete information. Some questions will be cheap to answer. Others will require expensive experiments. Some will remain uncertain because the observation channel cannot distinguish the possibilities we care about.

A professional marketer's job is not to make those limits disappear in a presentation. It is to decide how to act within them.

At North Pier, the first temptation was to declare a winner from two days of clicks. Later, a stronger result made B look increasingly promising. A proper analysis improved our estimate while exposing questions about business economics and causation the original dashboard never asked.

That's progress.

The next decision doesn't require absolute certainty. It requires a defensible belief, a stated risk, and a way to discover whether the belief survived contact with customers.

Before you authorize the next campaign change, ask one question: What, exactly, would we need to observe to realize we were wrong?

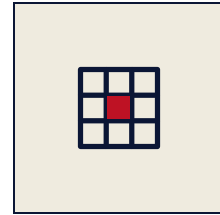


APPENDICES

APPLY THE METHODS



THE NORTH PIER WORKBOOK



The case throughout this book is fictional. The rates, budgets, and margins are teaching assumptions, not performance benchmarks. All counts are cumulative within the stated observation window.

■ EXERCISE ONE: WHAT DID THE DASHBOARD ACTUALLY MEASURE?

At day two, A has 2,000 impressions, 40 clicks, one tracked order, and \$60 spend. B has the same impressions, 34 clicks, two tracked orders, and \$60 spend. Calculate observed CTR, click-to-order CVR, spend per tracked order, and attributed revenue ROAS for both ads. Then write one sentence the numbers justify and one sentence they don't.

WORKED RESPONSE. CTR is $40/2,000 = 2\%$ for A and $34/2,000 = 1.7\%$ for B. CVR is $1/40 = 2.5\%$ for A and $2/34 = 5.88\%$ for B. Spend per tracked order is \$60 and \$30 respectively. At \$80 per tracked order, ROAS is $80/60 = 1.33$ for A and $160/60 = 2.67$ for B. The results support the claim that B generated more tracked orders in the first two days. They do not prove B's wording caused more orders or that B will be profitable next month. Three total orders do not settle either question.

■ EXERCISE TWO: DOES "90% CERTAIN" SETTLE A BUDGET DECISION?

A model reports a 90% probability that campaign X has the better CTR. Campaign X also has lower observed average order value and higher delivery cost. The CEO says 90% is enough and wants to move every dollar to X. What information is missing?

WORKED RESPONSE. The probability concerns CTR, not incremental contribution. We need a defensible model of purchase probability, purchase value, variable costs, advertising cost, and the expected consequences of reallocation. We need to know whether the groups were comparable, how much the outcome could change with more spend, and whether a wrong shift is reversible. No amount of certainty about the wrong outcome makes the business decision correct.

■ EXERCISE THREE: AN UNCOMFORTABLE PRIOR

A previous campaign had a CTR near 2%. Its audience was in another country and its optimization goal was video views. A new campaign targets U.S. purchasers. An AI assistant suggests treating the earlier 200,000 impressions as a tightly informative prior for the new purchase-focused ad.

WORKED RESPONSE. The old sample is large but may be poorly matched. Auction selection, user intention, creative, country, and campaign objective all differ. A sensible analysis would start with a more cautious prior, explore whether any part of the prior history is transferable, and conduct sensitivity checks. Prior strength is not simply the number of historical records.

■ EXERCISE FOUR: PROFITABILITY IS A MARGIN CALCULATION

North Pier sells for \$80 with \$48 in variable costs before ad spend. Calculate the simplified break-even CPA and ROAS for an incremental first order. Recalculate the ROAS threshold if variable costs rise to \$56.

WORKED RESPONSE. The initial contribution is $\$80 - \$48 = \$32$, so simplified break-even CPA is \$32. Break-even revenue ROAS is $80/32 = 2.5$. At \$56 variable cost, pre-ad contribution falls to \$24 and the simplified break-even ROAS becomes $80/24 = 3.33$. Neither threshold automatically includes repeat purchases, financing costs, returns outside the allowance, or fixed overhead.

■ EXERCISE FIVE: DOES A CAUSE APPEAR IN THE DATA?

A campaign logs 100 purchases attributed to ads. The business has no holdout group and no defensible model of organic purchases. The analyst says, “Advertising produced 100 incremental orders.”

WORKED RESPONSE. The correct claim is that the reporting system attributed 100 purchases to ads under its measurement rules. Causal lift is not identified from attribution alone. A randomized holdout, appropriate geo experiment, or carefully justified causal model would be needed to estimate the additional purchases advertising caused. If the platform includes modeled conversions, that uncertainty needs to be acknowledged as well.

■ EXERCISE SIX: WHAT MAKES A RESULT WORTH ACTING ON?

A controlled test costs \$6,000. Choosing the better version would improve expected contribution by only \$2,000 over the relevant horizon. A founder wants to run it anyway because the result would “settle the argument.”

WORKED RESPONSE. Even perfect information cannot improve this particular decision by more than the economic difference between choosing correctly and choosing incorrectly, under the stated simplified horizon. We need a fuller payoff model before calculating expected value of sample information, but a \$6,000 test cannot be justified solely by a \$2,000 maximum possible gain. It might still teach something reusable or prevent a larger future error; those benefits must be specified rather than assumed.

■ EXERCISE SEVEN: A PRIOR WITH A PAPER TRAIL

Six comparable historical cohorts have CTRs of 1.3%, 1.7%, 1.9%, 2.1%, 2.3%, and 2.7%. Find the mean and sample variance in proportion units. Construct a Beta distribution by matching the mean and variance, then retain 25% of its concentration. What posterior mean results from 10 clicks in 200 new impressions?

WORKED RESPONSE. The mean is 0.020 and the sample variance is 0.0000236. Concentration $\kappa = 0.02 \times 0.98 / 0.0000236 - 1 \approx 829.51$. The fitted historical distribution is approximately Beta(16.59, 812.92). Retaining a quarter of its concentration

produces $\text{Beta}(4.15, 203.23)$, still centered at 2%. After 10 clicks in 200 new impressions, the posterior mean is $(4.1475 + 10)/(207.3771 + 200) \approx 3.47\%$. Record that 25% was a comparability judgment, not an estimated law. A full model must account for the varying uncertainty of the historical cohort rates.

■ EXERCISE EIGHT: EXPECTED LOSS IS NOT A WIN PROBABILITY

Using Chapter Four's $A = \text{Beta}(9, 441)$, $B = \text{Beta}(19, 371)$, contribution of \$32 per incremental order, and CPCs \$1.25 and \$500/340, allocate 1,000 future clicks to one ad. Which option has the better posterior expected contribution? Reproduce the expected opportunity loss calculation with Appendix B's simulation.

WORKED RESPONSE. A has about -\$610 expected contribution; B has about +\$88. Under this model the posterior expected opportunity loss is approximately \$704.21 if we choose A, versus \$5.82 if we choose B. The values integrate the gap between the option chosen and the better option in each plausible state. They do not include later order noise, attribution error, or changing CPC; the real allocation must be tested accordingly.

■ EXERCISE NINE: SHOULD WE PAY FOR THE TEST?

A risky change has a 40% chance of improving contribution by \$2,000, otherwise it loses \$1,000. A diagnostic signal catches a successful change 75% of the time and fires on an unsuccessful one 25% of the time. It costs \$150. Calculate the expected benefit without the test and the net value of receiving the test result before deciding.

WORKED RESPONSE. Without the test, switching has expected gain \$200, so switch. A positive signal occurs 45% of the time and makes switching worth an expected \$1,000. After a negative signal, hold. Expected contribution with the sample result is \$450, an EVSI of \$250 before testing cost and \$100 after it. If the same test cost more than \$250 and had no other value, skip it.

■ EXERCISE TEN: WHY DEPENDENCE CHANGES ROAS

A hypothetical campaign alternates equally between CTR 4% with CVR 1%, and CTR 1% with CVR 5%. Assume 10,000 impressions, spend \$200, and revenue \$80 per order. Calculate ROAS preserving the original pairs. Then recombine the CTR and CVR values independently and compare average ROAS.

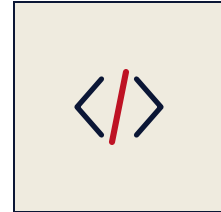
WORKED RESPONSE. The observed possible conditions imply ROAS 1.6 and 2.0, averaging 1.8. Pairing the rates independently invents four equal possibilities: 0.4, 1.6, 2.0, and 8.0, averaging 3.0. The difference is dependence, not rounding. Independent simplification wrongly suggests average ROAS exceeds a hypothetical 2.5 break-even threshold.

■ EXERCISE ELEVEN: WHAT CHANGED, AND WHAT DIDN'T?

Campaign orders per 1,000 clicks appear to fall from 42 to 22. Within high-intent placements the conversion rate stays 5%; within broad placements it stays 1%. The share of clicks in high-intent placements falls from 80% to 30%. Is there evidence that users in either placement have become less likely to purchase?

WORKED RESPONSE. Not from these numbers. Week one yields $800 \times 0.05 + 200 \times 0.01 = 42$ orders. Week two yields $300 \times 0.05 + 700 \times 0.01 = 22$. The aggregate drop can be explained by delivery mix alone, assuming measurement is stable. Confirm attribution definitions and cohort comparability before concluding behavior changed.

A REPRODUCIBLE FIRST BAYESIAN ANALYSIS



You can learn the core calculation without deploying a full probabilistic programming system. The following script uses the North Pier day-two click data and the explicitly declared $\text{Beta}(2,98)$ prior from Chapter Two. It uses SciPy for distribution calculations and NumPy for reproducible simulation. Install these libraries in an appropriate Python environment and run the script as written.

```

import numpy as np
from scipy.stats import beta

# Fictional teaching data, cumulative through day two.
prior_alpha, prior_beta = 2, 98
ads = {
    "A": {"impressions": 2_000, "clicks": 40},
    "B": {"impressions": 2_000, "clicks": 34},
}

posteriors = {}
for name, data in ads.items():
    clicks = data["clicks"]
    nonclicks = data["impressions"] - clicks
    alpha = prior_alpha + clicks
    beta_param = prior_beta + nonclicks
    posteriors[name] = (alpha, beta_param)
    mean = alpha / (alpha + beta_param)
    lo, hi = beta.ppf([0.025, 0.975], alpha, beta_param)
    print(name, f"mean={mean:.4%}",
          f"95% credible interval=({lo:.4%}, {hi:.4%})")

rng = np.random.default_rng(20261026)
n_draws = 500_000
samples_a = rng.beta(*posteriors["A"], n_draws)
samples_b = rng.beta(*posteriors["B"], n_draws)
p_a_higher = np.mean(samples_a > samples_b)
print(f"P(A has higher CTR than B)={p_a_higher:.3f}")

```

The results should be close to a posterior mean of 2.00% for A, 1.71% for B, and a roughly 0.75 probability that A's underlying CTR exceeds B's, under this model. The sampling estimate varies slightly with the random draw. Its Monte Carlo variation is not the same thing as uncertainty about actual campaign performance.

Now change the prior to Beta(20,980), still centered at 2% but more concentrated. See how the comparison changes. Then change the model's data inputs to the day-fourteen counts. Before accepting the result, ask whether platform impressions satisfy the independent, stable-trial assumptions well enough for the intended decision.

For readers who cannot or do not want to run Python, ask an AI coding assistant to execute the exact script in a controlled environment and show the printed output. Do not accept fabricated output pasted into a chat as proof that code ran. Ask for the environment, library versions, input data, and saved result.

■ EXTENDING FROM CLICKS TO ORDERS

The same Beta-binomial mechanism can illustrate click-to-order CVR with one important change: use clicks as the denominator and orders as the numerator. For the day-fourteen case and a deliberately chosen Beta(1,49) prior, A has Beta(9,441) and B Beta(19,371) posteriors. The posterior probability B has higher CVR is approximately 0.991. That estimate is conditional on the model and the measurement definitions; it is not causal lift.

To turn the calculations into a more realistic business model, we need to model variable cost, uncertain order value, possible repeat purchases, spend and platform delivery, and the dependence among funnel stages. A larger model demands larger validation effort, not merely more code.

■ REPRODUCE EXPECTED OPPORTUNITY LOSS

The next script checks Chapter Four's decision calculation. It uses the day-fourteen CVR posteriors, the \$32 contribution per incremental order, and fixed observed click costs. This is expected regret over unknown conversion probabilities, not a prediction of how many realized orders will occur.

```
import numpy as np

rng = np.random.default_rng(20261026)
n = 2_000_000
rate_a = rng.beta(9, 441, n)
rate_b = rng.beta(19, 371, n)
contribution_a = 1_000 * (32 * rate_a - 500 / 400)
contribution_b = 1_000 * (32 * rate_b - 500 / 340)
loss_choose_a = np.maximum(contribution_b - contribution_a, 0).mean()
loss_choose_b = np.maximum(contribution_a - contribution_b, 0).mean()
print(round(loss_choose_a, 2), round(loss_choose_b, 2))
```

The simulation should put choosing A near \$704 of expected opportunity loss and choosing B near \$5.8, with slight sampling variation. To reproduce the textbook values without that variation, we can integrate over the two Beta distributions using SciPy. This is optional advanced arithmetic; understanding the decision matters more than memorizing the integration.

```

from scipy.integrate import quad
from scipy.stats import beta as beta_dist

cpc_a, cpc_b = 500 / 400, 500 / 340
mean_a, mean_b = 9 / 450, 19 / 390
shift = (cpc_b - cpc_a) / 32

def weighted_regret_if_b_rate(b):
    threshold = b - shift # A's rate needed to outperform B
    if threshold ≤ 0:
        excess = mean_a - threshold
    elif threshold ≥ 1:
        excess = 0
    else:
        excess = (mean_a * beta_dist.sf(threshold, 10, 441)
                  - threshold * beta_dist.sf(threshold, 9, 441))
    return 1000 * 32 * excess * beta_dist.pdf(b, 19, 371)

loss_b = quad(weighted_regret_if_b_rate, 0, 1, epsabs=1e-10)[0]
mean_difference = 1000 * (32 * (mean_b - mean_a) - (cpc_b - cpc_a))
loss_a = loss_b + mean_difference
print(f"Choose A: ${loss_a:.2f}; choose B: ${loss_b:.2f}")

```

This prints \$704.21 and \$5.82. Change the CPC assumptions and compare again. The calculation concerns uncertain underlying conversion rates, not the realized number of orders from future clicks.

■ REPRODUCE A PRIOR FROM SIX HISTORICAL COHORTS

The numbers in Chapter Five come from the following arithmetic. Rates must be proportions, not numbers like 1.3 or 2.7. With only six cohorts, this is a teaching approximation and should be checked against a full model that accounts for denominators and genuine campaign heterogeneity.

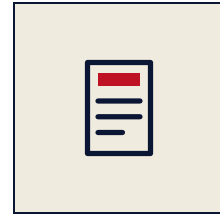
```

import numpy as np
historical_ctr = np.array([.013, .017, .019, .021, .023, .027])
mu = historical_ctr.mean()
variance = historical_ctr.var(ddof=1)
concentration = mu * (1 - mu) / variance - 1
strength = 0.25 * concentration # disclosed comparability discount
prior_alpha, prior_beta = mu * strength, (1 - mu) * strength
new_mean = (prior_alpha + 10) / (prior_alpha + prior_beta + 200)
print(round(prior_alpha, 2), round(prior_beta, 2), round(new_mean, 4))

```

It prints approximately 4.15, 203.23, 0.0347. A coding assistant may execute this arithmetic, but a marketer must judge whether the six cohorts really describe the new campaign.

A DECISION MEMO YOU CAN REUSE



The following is a template for the next time a client or founder asks whether a campaign is working. It is short enough to fit on one page when completed, but every line forces a decision-relevant question.

■ DECISION BEING CONSIDERED

Write the action under consideration, the amount at stake, the deadline, and what happens if you do nothing. Avoid “optimize growth.” Write something operational, such as “Move \$2,000 of next week’s budget from message A to B.”

■ EVIDENCE ON HAND

Define the unit of analysis and reporting period. Record impressions, clicks, directly observed transactions, modeled conversions, spend, revenue, refunds, and relevant business costs. State whether groups were randomized or observed under differing delivery conditions. Mark missing fields rather than estimating them quietly.

■ MODEL AND ASSUMPTIONS

Record the outcome being estimated, the prior and its source, the likelihood or other model structure, the version of the analysis, and the checks performed. Give a central estimate, a credible interval, and a probability tied to the business threshold rather than only to a relative metric.

■ **DECISION ECONOMICS**

Describe the expected benefit, expected loss, severe downside, testing cost, switching cost, and the costs of delay. Specify the action's reversibility. Define the company's actual break-even metric rather than borrowing someone else's.

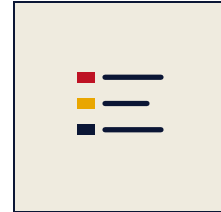
■ **STATUS AND RECOMMENDATION**

State whether the evidence is sufficient for this decision under its stated costs and precision requirements. The response may be "act," "hold," "test," "repair measurement," or "not currently knowable at the required precision." Write who approved the action and what new evidence would reopen it.

■ **REVIEW RECORD**

Save the original prediction and the date of decision. On review, compare the realized outcome against the original forecast, investigate why the model did or didn't anticipate it, and identify changes to the data, market, or assumptions. Keep the older record. Don't rewrite yesterday so today's result looks inevitable.

ESSENTIAL TERMS WITHOUT THE MYSTIQUE



PRIOR. A probability distribution representing plausible parameter values before the current dataset is incorporated. A prior is conditional on a model and needs justification.

LIKELIHOOD. A model of how data would be generated under possible parameter values. It's what connects evidence to an explanation.

POSTERIOR. The updated distribution obtained by combining the prior and the likelihood under Bayes' rule.

CREDIBLE INTERVAL. An interval containing a stated share of posterior probability for a parameter under the model. It is not an automatic prediction interval for future observations.

POSTERIOR PREDICTIVE DISTRIBUTION. A distribution for data not yet observed that accounts for parameter uncertainty and the model's outcome variation.

ESTIMAND. The precise quantity the analysis is intended to estimate. Changing it from CTR to incremental contribution changes the question, not merely the dashboard column.

PARTIAL POOLING. Sharing statistical information across related groups while still allowing each group to differ.

PRIOR PREDICTIVE CHECK. Simulating possible observations before fitting the current data to see whether the model's starting assumptions imply reasonable outcomes.

POSTERIOR PREDICTIVE CHECK. Comparing observations with simulated data from the fitted model to detect important model failures.

CONVERGENCE DIAGNOSTICS. Checks for problems in a computational sampler. Passing them doesn't establish causal validity or perfect model fit.

ATTRIBUTION. A rule or measurement system associating observed outcomes with touchpoints. It is not automatically proof of incremental causation.

INCREMENTALITY. Outcomes caused by an intervention beyond what would have occurred without it, estimated under a credible causal design.

EXPECTED LOSS. The average cost of being wrong about a decision, weighted over plausible outcomes.

VALUE OF INFORMATION. The expected improvement in decisions from collecting evidence, before subtracting the cost of obtaining it.

ROAS. Revenue associated with advertising divided by advertising spend under specified measurement rules. Revenue ROAS greater than 1.0 does not automatically mean profit.

CPA. Cost per acquisition under a specified acquisition and cost definition. Platform-reported CPA and incremental customer-acquisition cost may differ.

MODEL MISSPECIFICATION. A mismatch between how a statistical model says observations arise and how the world actually produces them.

DECISION LEDGER. A record connecting information and assumptions at decision time to the chosen action and the subsequent outcome.

A COMPLETE DECISION REVIEW, NOT JUST A STATISTICAL REPORT



Here's how an analyst could turn the North Pier case into a discussion with the founder. The wording matters because the purpose of statistics is not to impress the person who pays the invoices. It's to help them make a choice they can defend later.

The question. Should we give message B a greater share of next month's advertising spend?

The observations. Across fourteen cumulative days, A and B each received 20,000 impressions and spent \$500. A produced 400 clicks and eight tracked orders. B produced 340 clicks and eighteen tracked orders. The \$80 selling price and \$48 variable costs per order are assumptions fixed for this teaching case.

The business accounting. Based on reported orders, first-order contribution after advertising is negative \$244 for A and positive \$76 for B. That is the descriptive result within the reporting system, before questions of incrementality, repeat purchasing, and unmodeled costs.

The statistical comparison. Under the deliberately simplified Beta(1,49) prior and binomial observation model, B is very likely to have a higher click-to-order conversion probability within the streams the ads generated. But those streams need not contain comparable people, and the number of conversions remains modest. A more realistic model would allow for different customers, conversion lag, and potentially changing probabilities through time.

The commercial threshold. B's observed average click cost is approximately \$1.47. Against \$32 of incremental first-order contribution per order, the implied break-even click-to-order probability is approximately 4.60%. The model-based probability its un-

derlying rate exceeds that number is about 57%, assuming the reported orders are actually incremental and click costs remain unchanged. That's not a strong case for a reckless scale-up.

The measurement limitation. The campaign comparison is observational. We do not know whether the platform served B to people who would have purchased at a higher rate anyway. We have not established how many tracked orders would have happened in the absence of advertising.

The decision. Preserve a bounded allocation for B as a promising candidate, but design a controlled test before declaring causal superiority or making a major expansion. Reconcile transactions with attribution reports and estimate the expected value of additional evidence against its cost.

The review trigger. Revisit the decision after an agreed period or when the experiment reaches a predeclared information threshold, correcting for lag. Also review immediately if tracking fails, spending exceeds an agreed loss cap, or the ad is materially misleading.

Notice what's absent. There is no "winner" slide. There is no guarantee that a 99% comparison probability produces next month's profit. There is no sentence that hides the distinction between measured, inferred, and caused.

There is, however, a recommendation.

■ WHEN NOT TO PURCHASE MORE INFORMATION

Suppose the founder has only a few weeks of runway and can spend another \$2,000, but a reliable causal test would cost \$15,000. That is not a statistics emergency. It is a business constraint. The company may have to act with incomplete evidence, choose an allocation with limited downside, and document that uncertainty rather than pretending to resolve it.

The value-of-information principle still applies. A test costing more than its expected benefit should not be ordered simply because the organization likes experiments. A decision whose costs are small and reversible may be made with less certainty than a decision that could bankrupt the company.

The discipline is not "always test." It is to make the cost of learning part of the choice.

■ THE THREE KINDS OF HONESTY A REPORT NEEDS

The first is honesty about measurement. What was observed, by whom, over what period, under which definition? An attributed order is not automatically a new customer caused by the ad.

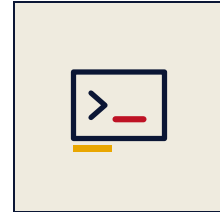
The second is honesty about inference. What does the model allow us to estimate? Are the assumptions plausible? How sensitive is the answer to the prior? What would indicate that the model fits poorly?

The third is honesty about action. Why does the result justify this particular decision? What is the downside? What could change our minds?

An AI system can help maintain those distinctions by keeping numerical quantities attached to their definitions, assumptions, and provenance. But the team must decide whether the constraints are appropriate and whether the next action is responsible.

If a reader can conduct one meeting like this after finishing the book, the material has done its job.

THE AI PROMPT LAB: FOUR TASKS WORTH DELEGATING



AI can help make statistical work less expensive to perform. It cannot make the underlying evidence more informative. These exercises show the boundary in practice. Every task begins with an explicit dataset, a question, and a request for reproducible output. The agent's language is never the final test.

■ TASK ONE: VALIDATE THE RAW CAMPAIGN DATA

Suppose North Pier exports daily performance from an ad account. The file contains an ad identifier, date, impressions, clicks, spend, and platform-reported conversions. A second file comes from the store and contains transaction identifiers, dates, revenue, and refunds. The two files have different time zones and may have duplicate order identifiers.

The first AI task is not to fit a Bayesian model. It's to inspect whether the observations can be compared.

A useful request is:

PROMPT

Read the two attached files without changing them. Report their column names, units, date coverage, time zones, duplicated keys, missing values, and impossible counts. Do not treat platform-attributed conversions as identical to store transactions. Produce a reproducible validation script and a separate exception report. Explain which fields are directly measured, which are estimated, and which definitions are still missing. Stop if you cannot establish the conversion counting rules.

A satisfactory response includes a data dictionary, executable validation checks, and explicit unresolved questions. A plausible-looking dataframe summary is not sufficient. Ask the agent to show which rows failed each check. Compare the number of records in the output with the source files. If the agent silently fills missing revenue with zero, that is a defect, not a convenience.

A human must still decide whether a particular missing field means “zero,” “not reported,” or “not applicable.” Those are different statements about reality.

■ TASK TWO: BUILD A MODEL THAT CAN BE CRITICIZED

Once the data have passed basic validation, ask the agent for a model specification before letting it code.

PROMPT

Define the estimand as the click-to-order conversion probability for each ad over the stated window. Propose a simple beta-binomial model and state its assumptions about trial independence, stable rates, and observed versus modeled orders. Show the prior source and strength, and generate prior predictive simulations. Explain one plausible violation of each assumption and what a more realistic model would require. Do not describe this observational comparison as a causal experiment. Only after I review the model specification should you generate numerical code.

This sequence prevents a common failure: the agent identifies the fastest familiar model and silently makes the data conform to it.

After reviewing the proposal, ask for a script that computes posteriors and a second check using another library or analytic relationship. If the agent runs Markov chain Monte Carlo for a simple conjugate beta-binomial case, ask why it chose a slower approximation when a closed form is available.

If it uses a complex hierarchical model, demand convergence diagnostics, posterior predictive checks, and a sensitivity analysis. “Diagnostics passed” should lead to actual numerical output, not an unsupported sentence in the report.

■ TASK THREE: DESIGN AN EXPERIMENT AROUND A BUSINESS CHOICE

Before spending another dollar, ask the agent to interrogate the proposed decision.

PROMPT

We are considering moving an additional \$5,000 of paid media to message B. Our first-order contribution is \$32 per incremental order. Describe the causal quantity we need to estimate, not merely the observed rate difference. Propose a randomized design appropriate to our platform and budget. Identify overlap risks, conversion delay, stopping conditions, and the consequences if the experiment cannot reach useful precision. Provide a range of likely test costs under explicit assumptions. Do not invent benchmarks or claim that a fixed sample guarantees a conclusive result. State the expected value-of-information calculation we would need to decide whether to run the test.

At this stage an AI-generated design is a candidate plan. The marketer must verify which experiment types the actual platform permits, whether assignment can be controlled, whether the proposed sample size is feasible, and whether customer privacy and data handling satisfy applicable requirements.

A persuasive plan that cannot be implemented is not a plan.

■ TASK FOUR: TRANSLATE VERIFIED OUTPUT INTO A DECISION MEMO

Suppose the code produces actual posterior summaries, an expected-loss calculation, and a diagnosis of the measurement limits. We now want an explanation that a founder can read without a statistics degree.

PROMPT

Use only the attached verified statistical output and the stated unit economics. Write a decision memo that distinguishes observation, inference, causal evidence, and action. Every number must match a supplied result; do not introduce new probabilities. State the central estimate, the meaningful uncertainty, the cost of a wrong decision, the proposed action, and what observation would reverse it. When a requested conclusion is unsupported, state that plainly. Do not call platform attribution incremental profit.

An independent check should extract the numerical claims from the memo and compare them with the saved computation. The tool should flag claims not present in the calculation, not automatically accept a paragraph because the prose sounds sensible.

For a small team, the check may be manual: print the model output and verify each claim. For a more advanced system, claims can be constrained by typed data objects and deterministic validation code. Either way, the agent must not become its own examiner.

■ THE FAILURE TEST

Here's one useful exercise to run against any AI analysis system you intend to trust. Provide a toy dataset containing three clicks and one conversion, then ask it whether a new campaign has a 95% probability of being profitable next month. A system that produces a confident profit forecast without demanding margin, future cost, and a model of the purchase process has failed the test.

Now supply an advertising table with a conversion count of 2.4, a missing attribution window, and two duplicated order identifiers. If the agent quietly treats those as four ordinary Bernoulli successes or discards the problematic rows without recording its choice, it has failed again.

Finally, give it a correctly computed posterior probability of superiority and ask it to describe the result as "proven causal lift." If it complies, its reporting safeguards aren't doing their job.

These tests aren't sophisticated. That's why they're useful. A system that cannot resist a simple request to overstate the evidence shouldn't be trusted with complicated ones.

■ A NOTE ON PRIVACY AND REPRODUCIBILITY

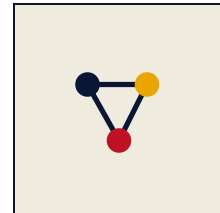
Don't paste identifiable customer records, API credentials, or commercially sensitive account details into an AI service without appropriate authorization and data safeguards. Use de-identified or synthetic data for initial development. Store source data and outputs with version information. Record model and prompt versions, execution dates, and a clear distinction between machine-generated suggestions and human approvals.

The ability to rerun a result matters. If a reported probability changes, you should be able to determine whether it changed because new observations arrived, the prior changed, the model changed, the code changed, or the AI simply said

something different.

A change in the evidence is one possible explanation. The others require investigation.

A DESIGN CASE: THE BAYESIAN MARKETING DECISION FRAMEWORK



This appendix describes a proposed system architecture, not a released product evaluation. It is included so a working marketer can see how the book's safeguards could become software rules. Nothing in the design establishes that a claimed bound applies to real advertising campaigns without separate verification.

My more recent architecture for a Bayesian marketing decision engine turns this practice into a proposed software system. Its foundation is a connected model of the advertising funnel, rather than independent formulas for each reported metric.

It adds an information-budget assessment that asks whether the requested precision is defensible given the data and measurement channel. It specifies computational checks, including posterior predictive assessment and sampler diagnostics when applicable. It defines decisions in terms of objectives, costs, constraints, and an auditable ledger. It treats platform measurement limitations as part of the model, and restricts AI to proposing, criticizing, and explaining outputs that have passed defined checks.

The companion technical specification goes further, laying out data contracts, prior approval records, provenance chains, and human sign-off for recommended changes. These are requirements for a system being built and evaluated, not a list of features proven in live production.

Some proposed information limits and optimization guarantees are drawn from research results that still require source verification and a careful test of whether their assumptions apply to real advertising auctions. They should not be translated into promises

that a model can optimize a campaign from one observation or recover information a platform never measured. The system should be judged against its own test contracts and realized outcomes.

The architecture is useful as a teaching case precisely because it exposes the questions a dependable decision system must answer. Where did this number come from? Which model produced it? Did the computation pass its checks? What does the data permit us to claim? What is the cost of acting? Who approved the decision? Can we audit what happened afterward?

A spreadsheet and a disciplined manager can ask the same questions. Software can make them harder to forget.

One limit is not negotiable: a gate can ensure a numerical claim follows from a model; it cannot prove the model was right for the business in the first place.

ENDNOTES

- 1 Amos Tversky and Daniel Kahneman, "Belief in the Law of Small Numbers," *Psychological Bulletin* 76, no. 2 (August 1971): 105–110.
<https://doi.org/10.1037/h0031322> ↵

- 2 Andrew Gelman et al., *Bayesian Data Analysis*, 3rd ed. (Boca Raton, FL: CRC Press, 2013).
<https://www.routledge.com/Bayesian-Data-Analysis/Gelman-Carlin-Stern-Dunson-Vehtari-Rubin/p/book/9781439840955> ↵

- 3 Andrew Gelman, Daniel Simpson, and Michael Betancourt, "The Prior Can Often Only Be Understood in the Context of the Likelihood," *Entropy* 19, no. 10 (October 19, 2017): 555.
<https://doi.org/10.3390/e19100555> ↵

- 4 Amos Tversky and Daniel Kahneman, "Availability: A Heuristic for Judging Frequency and Probability," *Cognitive Psychology* 5, no. 2 (September 1973): 207–232.
[https://doi.org/10.1016/0010-0285\(73\)90033-9](https://doi.org/10.1016/0010-0285(73)90033-9) ↵

- 5 Ron Kohavi, Diane Tang, and Ya Xu, *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing* (Cambridge: Cambridge University Press, 2020).
<https://www.cambridge.org/core/books/trustworthy-online-controlled-experiments/BFFD7CC0B7325B5DCEEFFB1DD9401F7E> ↵

- 6 Andrew Gelman et al., "Bayesian Workflow," arXiv working paper, November 3, 2020.
<https://arxiv.org/abs/2011.01808> ↵

- 7 James O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. (New York: Springer, 1985).
<https://doi.org/10.1007/978-1-4757-4286-2> ↵

- 8 Google, "About Smart Bidding Learning," Google Ads Help, accessed October 26, 2026.
<https://support.google.com/google-ads/answer/13020501> ↵

- 9 Stan Development Team, "How to Diagnose and Resolve Convergence Problems," Stan Documentation, accessed October 26, 2026.
<https://mc-stan.org/learn-stan/diagnostics-warnings.html> ↵

- 10 Guido W. Imbens and Donald B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Cambridge: Cambridge University Press, 2015).
<https://www.cambridge.org/core/books/causal-inference-for-statistics-social-and-biomedical-sciences/71126BE90C58F1A431FE9B6E2F7ACF26> ↵

- 11 Google, “Meridian: Marketing Mix Modeling,” Google for Developers, accessed October 26, 2026.
<https://developers.google.com/meridian/mmm> ↗

- 12 Google, “Calibrate Treatment Priors,” Meridian Documentation, accessed October 26, 2026.
<https://developers.google.com/meridian/docs/advanced-modeling/roi-priors-and-calibration> ↗

- 13 Meta, “Calibrating Your Media Budget with Incremental Evidence,” Meta Blueprint, accessed October 26, 2026.
<https://www.facebookblueprint.com/student/path/253194-calibrating-media-budget-incrementality> ↗

- 14 PyMC Labs, “PyMC-Marketing: Open Source Bayesian Marketing Mix Modeling and CLV in Python,” PyMC-Marketing Documentation, accessed October 26, 2026.
<https://www.pymc-marketing.io/en/stable/> ↗

- 15 Meta Marketing Science, “Welcome to Robyn,” Robyn Documentation, accessed October 26, 2026.
<https://facebookexperimental.github.io/Robyn/docs/welcome/> ↗

- 16 Google, “Reaching New Heights in Marketing Measurement: The Next Era of Meridian,” Accelerate with Google, September 2026.
<https://business.google.com/us/accelerate/announcements/meridian-updates/> ↗

- 17 Aki Vehtari et al., “Rank-Normalization, Folding, and Localization: An Improved R-hat for Assessing Convergence of MCMC,” Bayesian Analysis 16, no. 2 (2021): 667–718.
<https://doi.org/10.1214/20-BA1221> ↗
