MarketDial

White Paper Series

Accurate decision intelligence

Achieving high confidence:

Why patience in reading test results is key

Ally Mauldin, Data Scientist, MarketDial

Contents

Introduction

Many retailers and businesses who conduct A/B testing at their physical sites wonder how long they need to run a test before they can trust the results and make a decision. Will four weeks be enough time? Sixteen weeks?

While it can be exciting to see results straight out of the gate, more often than not, test results are not reliable for at least 4 - 5 weeks post launch. As *Harvard Business Review* notes, "too many managers don't let tests run their course. Because most of the software for running these tests lets you watch results in real time, managers want to make decisions too quickly. . . . The problem is that, because of randomization, it's possible that if you let the test run to its natural end, you might get a different result."

To achieve high confidence, or statistical significance, most tests should run for multiple weeks, and even for tests that achieve high confidence the first week, early test results still may not reflect final outcomes. Understanding why and how this occurs can help business leaders tame overeagerness in exchange for reliability.

Raising confidence

Reaching statistical significance or achieving high confidence in A/B testing takes time; the wait helps ensure that the results are not due to random chance but are a true reflection of consumer behavior in response to change.

Confidence is expressed in a percentage that helps reflect the reliability of test outcomes. As a probability measurement, the higher the confidence, the lower the risk:

- **50% confidence or lower** – unreliable, might as well flip a coin, high risk
- **51% to 80% confidence** – good indication of direction (positive or negative results) but not a good prediction of exact numbers, moderate risk
- **80% or higher** – reliable data, low risk
- **90% to 95%** – ideal, particularly for high-investment initiatives

To reach high confidence, time and patience are essential. The duration of an A/B test is crucial for enhancing the reliability of its results because it allows for a larger volume of data collection, reducing the impact of random fluctuations and external factors.

Early tests often do not have enough data points to lead to high confidence. With more test sites or more weeks of data, it becomes easier to determine whether the lift from a test is due to random chance or not.

High confidence supports a culture of data-driven decision making within an organization, providing a framework for continually testing, learning, and adapting based on reliable empirical evidence rather than intuition or bias.

Getting the test length right

How much longer?

The best approach for in-store testing is to run a test for at least **4 – 5 weeks** (and ideally even longer) to make sure you collect enough data and see steady results before making a decision.

While it's important to run a test for a long enough period before using the test results to make a decision, how long is long enough?

There's a very natural inclination to want to check in on how a test is doing as soon as it has kicked off – and to make a decision after only letting the test run for a week or two. But most tests need to run for multiple weeks before reaching statistical significance.

Test length stabilizes outcomes and improves confidence by ensuring enough data is collected for statistical significance. Longer tests gather larger sample sizes, reducing random variations and accounting for seasonality or external factors. This increases statistical power, minimizing the likelihood of false positives and negatives and narrowing confidence intervals for more precise estimates. Consequently, longer test lengths provide more reliable data, leading to better-informed decisions and improved outcomes.

Case in point

Shifting employee hours

Situation

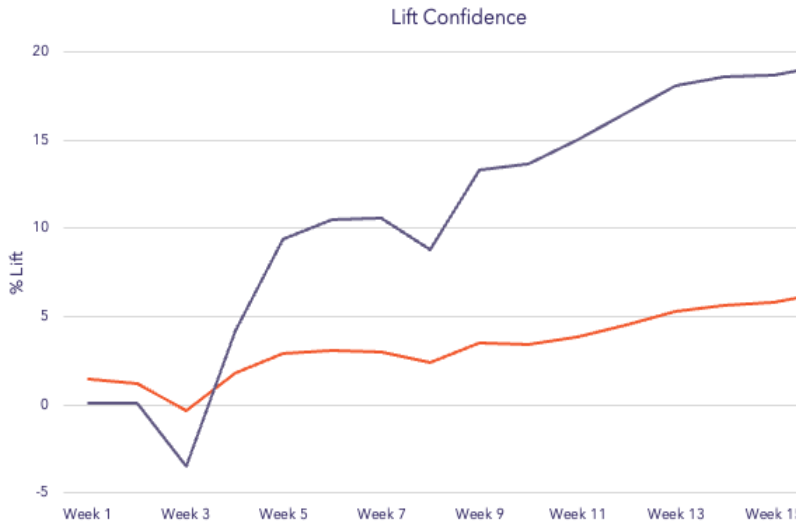
A fashion retailer wanted to see if increasing employee hours during their busiest times on the weekend would lead to an increase in revenue.

Solution

They created a test comparing stores with the increase in hours against stores without the increase.

Result

Since the test could only be reasonably measured during the weekends, it took awhile for the true impact to emerge from the data. The first weeks of the test showed low confidence, but it steadily increased over the course of the test. After twelve weeks, results were clear: increasing employe hours led to a signifcant increase in revenue.



Key Takeaways

- Patience was pivotal to understanding the true impact of the test.
- Directional results were achieved after five weeks of testing and were clear by twelve weeks.

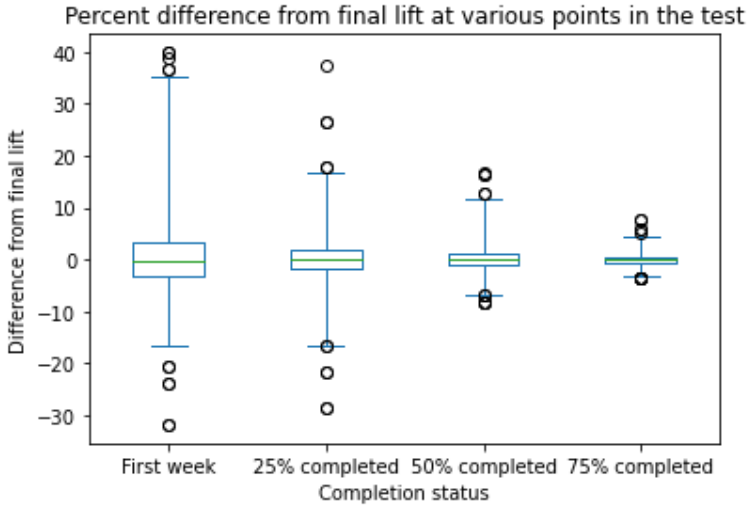
Early results vs. end results

Even for tests that achieve high statistical confidence in the first week, MarketDial research shows those results are not always indicative of final test outcomes.

Tests with high confidence early on can still vary greatly from the beginning of the test to the end. This shift in results can happen across all types of tests. In fact, we looked at all tests built in MarketDial that had 90% confidence or higher in the first week and analyzed how the lift changed by the end of the test. We found that:

- **30% of tests had lift change by more than five percentage points**
- **14% of tests ended with directionally different results**

This research underscores the importance of running tests for their full duration, even when early results seem promising. The variability observed in the lift changes and the occurrence of directionally different outcomes highlight that early confidence is not always a reliable predictor of final results. To make well-informed decisions, it's crucial to allow tests to run their course, ensuring that the insights gained are both accurate and actionable.



Case in point

Increased menu pricing

Situation

A quick service restaurant decided to increase prices in their stores that were located in more premium neighborhoods.

Solution

They created a test comparing stores with the increase in pricing against stores without the increase.

Result

The initial read of the test showed a success – revenue was up and transactions were flat; however, after a few weeks, transactions started to drop and revenue soon followed.

Key Takeaways

- Some tests may show early positive results, but by the end of the test period, the outcomes are negative.
- Such an outcome is more common in pricing tests because customers who discover a price increase will often complete the initial purchase but are less likely to return.

Case in point

Vetting CapEx investments

Situation

A convenience store installed new fountain drink dispensers and wanted to assess the return on investment.

Solution

They created a test comparing stores with the new dispensers against stores without the new dispensers.

Result

The initial read of the test showed was strongly negative; however, after two weeks, the results started to trend upward, and by the end of the test were strongly positive. The shift was likely due to either an adjustment period by the cusotmers or a longer installation period than initially anticipated.

Key Takeaways

- Some tests start with negative lift but turn around and are positive by the end of the test.
- When a big change is made, it's important to wait to read the results to account for unpredictable variables such as extended implementation or customer habituation.

Worth the wait

Reaching statistical significance in A/B testing is a critical milestone that signifies the reliability of the test results, enabling businesses to predict outcomes with greater accuracy.

When A/B test results achieve statistical significance, it means there is a sufficiently low probability that the observed differences between the test groups occurred by chance. This level of certainty is essential for predicting the actual impact of changes before they are rolled out on a larger scale.

Armed with statistically significant results, decision-makers can take action knowing that their choices are supported by solid data, not conjecture or temporary market conditions. This predictive power reduces the risk associated with new initiatives and ensures that resources are invested in strategies most likely to succeed. It also encourages a culture of testing and evidence-based decision making within the organization, fostering a proactive approach to business strategy.

Confidence matters

Ultimately, the ability to reach statistical significance in A/B testing not only empowers you to make better decisions but also drives continuous improvement and innovation, keeping you competitive and responsive to customer needs and market dynamics.

Why MarketDial?

MarketDial delivers decision intelligence by automating the data science behind in-store A/B testing. Retailers can now accurately measure ROI, enabling them to know on a small scale what changes will have big-scale impacts before full-scale rollout. For any retailer asking, "*What if,*" MarketDial provides the answers.

Ally Mauldin
Data Scientist, MarketDial

Ally Mauldin graduated from Utah State University with a bachelor's degree in math and statistics education and a minor in computer science. For three years her primary goal was to help teenagers at a charter school discover how math can be fun. She brought that passion and love for teaching with her when she joined the MarketDial team, where she now teaches clients how to make testing fun. Outside of work, you can find Ally mountain biking the Green Canyon trail in Logan or spending time with her husband and son. She also has been known to hold her own on go-kart race tracks.

© MarketDial 2024