3

Twyman’s Law and Experimentation Trustworthiness

> Twyman’s law, perhaps the most important single law in the whole of data analysis… The more unusual or interesting the data, the more likely they are to have been the result of an error of one kind or another
> *– Catherine Marsh and Jane Elliott (2009)*

> Twyman’s Law: “Any figure that looks interesting or different is usually wrong”
> *– A.S.C. Ehrenberg (1975)*

> Twyman’s Law: “Any statistic that appears interesting is almost certainly a mistake”
> *– Paul Dickson (1999)*

William Anthony Twyman was a UK radio and television audience measurement veteran (MR Web 2014) credited with formulating Twyman’s law, although he apparently never explicitly put it in writing, and multiple variants of it exist, as shown in the above quotations.

When we see a surprisingly positive result, such as a significant improvement to a key metric, the inclination is to build a story around it, share it, and celebrate. When the result is surprisingly negative, the inclination is to find some limitation of the study or a minor flaw and dismiss it.

Experience tells us that many extreme results are more likely to be the result of an error in instrumentation (e.g., logging), loss of data (or duplication of data), or a computational error.

To increase trust in experiment results, we recommend a set of tests and practices to indicate that something may be wrong with the results. In databases, there are integrity constraints; in defensive programming, we are encouraged to write `assert()`s to validate that constraints hold. In experimentation, we can run tests that check for underlying issues, similar to asserts:

if every user should see either Control or Treatment from a certain time, then having many users in both variants is a red flag; if the experiment design calls for equal percentages in the two variants, then large deviations that are probabilistically unlikely should likewise raise questions. Next, we share some great examples of findings that fit Twyman's law, and then discuss what you can do to improve the trustworthiness of controlled experiments.

Misinterpretation of the Statistical Results

Here are several common errors in interpreting the statistics behind controlled experiments.

Lack of Statistical Power

In our framework of Null Hypothesis Significance Testing (NHST), we typically assume that there is no difference in metric value between Control and Treatment (the Null hypothesis) and reject the hypothesis if the data presents strong evidence against it. A common mistake is to assume that just because a metric is not statistically significant, there is no Treatment effect. It could very well be that the experiment is underpowered to detect the effect size we are seeing, that is, there are not enough users in the test. For example, an evaluation of 115 A/B tests at GoodUI.org suggests that most were underpowered (Georgiev 2018). This is one reason that it is important to define what is practically significant in your setting (see Chapter 2) and ensure that you have sufficient power to detect a change of that magnitude or smaller.

If an experiment impacts only a small subset of the population, it is important to analyze just the impacted subset; even a large effect on a small set of users could be diluted and not be detectable overall (see Chapter 20 and Lu and Liu (2014)).

Misinterpreting p-values

P-value is often misinterpreted. The most common interpretation error is the belief that the p-value represents the probability that the average metric value in Control is different from the average metric value in Treatment, based on data in a single experiment.

The p-value is the probability of obtaining a result equal to or more extreme than what was observed, assuming that the Null hypothesis is true. The conditioning on the Null hypothesis is critical.

Here are some incorrect statements and explanations from *A Dirty Dozen: Twelve P-Value Misconceptions* (Goodman 2008):

1. *If the p-value = .05, the Null hypothesis has only a 5% chance of being true.* The p-value is calculated assuming that the Null hypothesis is true.
2. *A non-significant difference (e.g., p-value >.05) means there is no difference between groups.*

   The observed results are consistent with the Null hypothesis of zero Treatment effect and a range of other values. When confidence intervals are shown for a typical controlled experiment, then it includes zero. This does not mean that zero is more likely than other values in the confidence interval. It could very well be that the experiment is under-powered.
3. *P-value = .05 means that we observed data that would occur only 5% of the time under the Null hypothesis.*

   This is incorrect by the definition of p-value above, which includes equal or more extreme values than what was observed.
4. *p-value = .05 means that if you reject the Null hypothesis, the probability of a false positive is only 5%.*

   This is like the first example, but harder to see. The following example might help: Suppose you are trying to transmute lead to gold by subjecting the lead to heat and pressure and pouring elixirs on it. You measure the amount of "goldliness" in the resulting concoction, a noisy measurement. Since we know that chemical Treatments can't change the atomic number of lead from 82 to 79, any rejection of the Null hypothesis (of no change) would be false, so 100% of rejections are false positives, regardless of the p-value. To compute the false positive rate, that is, when the p-value is < 0.05 and yet the Null Hypothesis is true (note, conjunction, not conditioned on the Null hypothesis being true), we could use Bayes Theorem and would require some prior probability.

Even the above common definition of p-value, which assumes that the Null hypothesis is true, is not explicitly stating other assumptions explicitly, such as how the data was collected (e.g., randomly sampled) and what assumptions the statistical tests make. If an intermediate analysis was done, which impacted the choice of analysis to present, or if a p-value was selected for presentation because of its small size, then these assumptions are clearly violated (Greenland et al. 2016).

Peeking at p-values

When running an online controlled experiment, you could continuously monitor the p-values. In fact, early versions of the commercial product Optimizely encouraged this (Johari et al. 2017). Such multiple hypothesis testing results in significant bias (by 5−10x) in declaring results to be statistically significant. Here are two alternatives:

1. Use sequential tests with always valid p-values, as suggested by Johari et al. (2017), or a Bayesian testing framework (Deng, Lu and Chen 2016).
2. Use a predetermined experiment duration, such as a week, for the determining statistical significance.

Optimizely implemented a solution based on the first method, whereas the experimentation platforms being used at Google, LinkedIn, and Microsoft use the second.

Multiple Hypothesis Tests

The following story comes from the fun book, *What is a p-value anyway?* (Vickers 2009):

Statistician: Oh, so you have already calculated the p-value?
Surgeon: Yes, I used multinomial logistic regression.
Statistician: Really? How did you come up with that?
Surgeon: I tried each analysis on the statistical software drop-down menus, and that was the one that gave the smallest p-value.

The multiple comparisons problem (Wikipedia contributors, Multiple Comparisons problem 2019) is a generalization of peeking described above. When there are multiple tests, and we choose the lowest p-value, our estimates of the p-value and the effect size are likely to be biased. This is manifested in the following:

1. Looking at multiple metrics.
2. Looking at p-values across time (peeking as noted above).
3. Looking at segments of the population (e.g., countries, browser type, heavy/light, new/tenured).
4. Looking at multiple iterations of an experiment. For example, if the experiment truly does nothing (an A/A), running it 20 times may result in a p-value smaller than 0.05 by chance.

False Discovery Rate (Hochberg and Benjamini 1995) is a key concept to deal with multiple tests (see also Chapter 17).

Confidence Intervals

Confidence intervals, loosely speaking, quantify the degree of uncertainty in the Treatment effect. The confidence level represents how often the confidence interval should contain the true Treatment effect. There is a duality between p-values and confidence intervals. For the Null hypothesis of no-difference commonly used in controlled experiments, a 95% confidence interval of the Treatment effect that does not cross zero implies that the p-value is < 0.05.

A common mistake is to look at the confidence intervals separately for the Control and Treatment, and assume that if they overlap, the Treatment effect is not statistically different. That is incorrect, as shown in Statistical Rules of Thumb (van Belle 2008, section 2.6). Confidence intervals can overlap as much as 29% and yet the delta will be statistically significant. The opposite, however, is true: if the 95% confidence intervals do not overlap, then the Treatment effect is statistically significant with p-value < 0.05.

Another common misunderstanding about confidence intervals is the belief that the presented 95% confidence interval has a 95% chance of containing the true Treatment effect. For a specific confidence interval, the true Treatment effect is either 100% within it, or 0%. The 95% refers to how often the 95% confidence intervals computed from many studies would contain the true Treatment effect (Greenland et al. 2016); see Chapter 17 for more details.

Threats to Internal Validity

Internal validity refers to the correctness of the experimental results without attempting to generalize to other populations or time periods. Here are some common threats:

Violations of SUTVA

In the analysis of controlled experiments, it is common to apply the Stable Unit Treatment Value Assumption (SUTVA) (Imbens and Rubin 2015), which states that experiment units (e.g., users) do not interfere with one another. Their behavior is impacted by their own variant assignment, and not by the assignment of others. The assumption could clearly be violated in settings, including the following:

- Social networks, where a feature might spillover to a user's network.
- Skype (a communication tool), where peer-to-peer calls can violate SUTVA.

- Document authoring tools (e.g., Microsoft Office and Google Docs) with co-authoring support.
- Two-sided marketplaces (such as ad auctions, Airbnb, eBay, Lift, or Uber) can violate SUTVA through the "other" side. For example, lowering prices for Treatment has impact on Controls during auctions.
- Shared resources (such as CPU, storage, and caches) can impact SUTVA (Kohavi and Longbotham 2010). If the Treatment leaks memory and causes processes to slow down due to garbage collection and possibly swapping of resources to disk, all variants suffer. In an experiment we ran, the Treatment crashed the machine in certain scenarios. Those crashes also took down users who were in Control, so the delta on key metrics was not different—both populations suffered similarly.

See Chapter 22 for ways to address some of these violations.

Survivorship Bias

Analyzing users who have been active for some time (e.g., two months) introduces survivorship bias. A great example of this problem and the biases it introduces comes from World War II, when there was a decision to add armor to bombers. Recordings were made about where the planes took the most damage, and the military naturally wanted to add armor where the planes were hit the most. Abraham Wald pointed out that these were the **worst** places to add armor. Bullet holes were almost uniformly distributed, so armor should be added to the places where there were no bullet holes because bombers that were hit in those places... never made it back to be inspected (Denrell 2005, Dmitriev, et al. 2016).

Intention-to-Treat

In some experiments, there is non-random attrition from the variants. For example, in medical settings, patients in a Treatment may stop taking a medication if it has side effects. In the online world, you may offer all advertisers the opportunity to optimize their ad campaign, but only some advertisers choose to do the suggested optimization. Analyzing only those who participate, results in selection bias and commonly overstates the Treatment effect. Intention-to-treat uses the initial assignment, whether it was executed or not. The Treatment effect we are measuring is therefore based on the offer, or intention to treat, not whether it was actually applied.

In display advertising and e-mail marketing, we do not observe the Control group exposure and there are techniques proposed to address this motivated by intent-to-treat (Barajas et al. 2016).

Sample Ratio Mismatch (SRM)

If the ratio of users (or any randomization unit) between the variants is not close to the designed ratio, the experiment suffers from a Sample Ratio Mismatch (SRM). For example, if the experiment design is for a ratio of one-to-one (equally sized Control and Treatment), then deviations in the actual ratio of users in an experiment likely indicate a problem (see Chapter 21) that requires debugging. We share some examples below.

With large numbers, a ratio smaller than 0.99 or larger than 1.01 for a design that called for 1.0 more than likely indicates a serious issue. The experimentation system should generate a strong warning and hide any scorecards and reports, if the p-value for the ratio is low (e.g., below 0.001).

As defined earlier, the p-value is the probability of obtaining a result equal to or more extreme than what was observed, assuming that the Null hypothesis is true. If the experiment design was for equal allocations to both variants, then by design you should get a ratio close to 1.0, that is, the Null hypothesis *should be* true. The p-value thus represents the probability that the ratio we observed, or more extreme, is consistent with our experimentation system's design. This simple test has identified numerous issues in experiments, many that looked either great or terrible initially and invoked Twyman's law. Here are some other examples:

- **Browser redirects** (Kohavi and Longbotham 2010).

  A very common and practical mechanism to implement an A/B test is to redirect the Treatment to another page. Like many ideas, it is simple, elegant, and wrong; several different attempts have shown that this consistently causes an SRM. There are several reasons:

  a. **Performance differences.** Users in the Treatment group suffer an extra redirect, which may appear fast in the lab, but delays for users may be significant, on the order of hundreds of milliseconds, which has significant impact on key metrics (see Chapter 5).

  b. **Bots.** Robots handle redirects differently: some may not redirect on the http-equiv="REFRESH" meta tag; some will tag this as a new page worthy of deep crawling and crawl it more often.

  c. **Redirects are asymmetric.** When users are redirected to the Treatment page, they may bookmark it or pass a link to their friends. In most implementations, the Treatment page does not check that the user should really have been randomized into the Treatment, so this causes contamination.

The lesson here is to avoid redirects in implementations and prefer a server-side mechanism. When that is not possible, make sure that both Control and Treatment have the same "penalty," that is, redirect both the Control and Treatment.

- **Lossy instrumentation** (Kohavi and Longbotham 2010, Kohavi, Messner et al. 2010, Kohavi et al. 2012, Zhao et al. 2016)

  Click tracking is typically done using web beacons (typically a 1x1 GIF sent to the server to signal a click), which is known to be lossy (i.e., not 100% of clicks are properly recorded). This is not normally an issue, as the loss is similar for all variants, but sometimes the Treatment can impact the loss rate, making low-activity users (e.g., those who only had a single click) appear at a different rate and cause an SRM. When the web beacon is placed in a different area of the page, timing differences will skew the instrumentation.

- **Residual or carryover effects**

  New experiments usually involve new code and the bug rate tends to be higher. It is common for a new experiment to cause some unexpected egregious issue and be aborted or kept running for a quick bug fix. After the bug is fixed, the experiment continues, but some users were already impacted. In some cases, that residual effect could be severe and last for months (Kohavi et al. 2012, Lu and Liu 2014). This is why it is important to run pre-experiment A/A tests (see Chapter 19) and proactively re-randomize users, recognizing that in some cases the re-randomization breaks the user consistency, as some users bounce from one variant to another.

  The opposite could also be true. At LinkedIn, a new version of the People You May Know algorithm was evaluated and turned out to be highly beneficial, increasing user visits. When the experiment was stopped and restarted, there was a significant carryover effect from the prior experiment, enough to create an SRM and invalidate the results (Chen, Liu and Xu 2019).

  Residual information in browser cookies can impact experiments. Take, for example, an educational campaign that shows a message to users in Treatment, but in order to avoid bothering users, the message is shown only three times. The implementation uses a browser cookie that counts the number of times the message was shown. If the experiment is restarted, some Treatment users will have the cookie with a count > 0, and thus will either see fewer impressions or none at all, diluting the Treatment effect or creating an SRM (Chen et al. 2019).

- **Bad hash function for randomization**
  Zhao et al. (2016) describe how Treatment assignment was done at Yahoo! using the Fowler-Noll-Vo hash function, which sufficed for single-layer randomization, but which failed to properly distribute users in multiple concurrent experiments when the system was generalized to overlapping experiments. Cryptographic hash functions like MD5 are good (Kohavi et al. 2009) but slow; a non-cryptographic function used at Microsoft is Jenkins SpookyHash (www.burtleburtle.net/bob/hash/spooky.html).
- **Triggering impacted by Treatment**
  It is common to only trigger a segment of users into an experiment. For example, you may only trigger users in a certain country, say the US. These users are then randomly split into the variants.

  If triggering is done based on attributes that are changing over time, then you must ensure that no attributes used for triggering could be impacted by the Treatment. For example, assume you run an e-mail campaign that triggers for users who have been inactive for three months. If the campaign is effective, those users become active and the next iteration of the campaign could have an SRM.
- **Time-of-Day Effects**
  Let's demonstrate this again using an e-mail campaign setup as an A/B test with different e-mail body text for each variant. In the real example, users were properly randomized into equally sized Control and Treatment groups, yet the e-mail open rates, which should be approximately the same, showed up as an SRM.

  A long investigation found that the open times clustered around different time periods, which led to the conjecture, later confirmed, that due to ease of implementation, the e-mails were first sent to Control users and then to Treatment users—the first group received the e-mails during work hours, whereas the second group received them after work.
- **Data pipeline impacted by Treatment.**
  The MSN portal (www.msn.com) has an Info Pane area on the page with multiple "slides" that rotate and a dot that indicates each slide (see arrow on Figure 3.1) (Kohavi 2016).

  A key component of the MSN OEC is clicks-per-user, which represents user engagement. The team ran an experiment where the Treatment increased the number of slides in the Info Pane from 12 to 16.

  Initial results showed a significant reduction in user engagement for the Treatment, but the experiment had an SRM: the ratio was 0.992 instead of 1.0. With over 800,000 users in each variant, the p-value of such a split was 0.0000007, which meant that the probability of such a split happening by

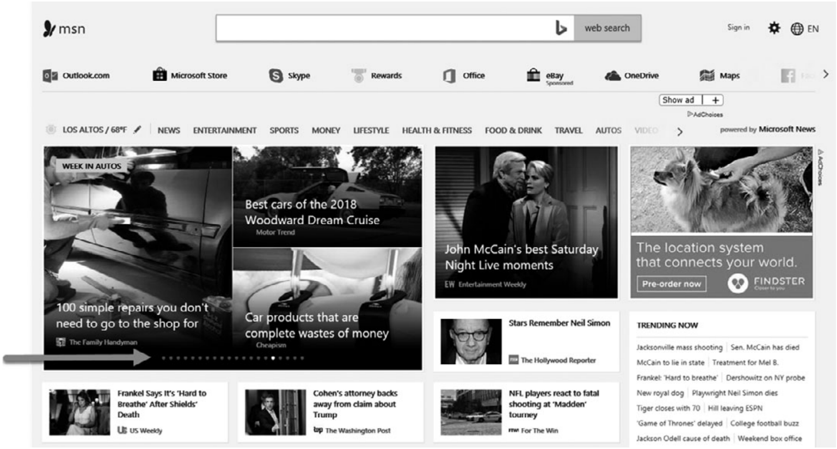
msn
web search
Sign in
EN
Outlook.com
Microsoft Store
Skype
Rewards
Office
eBay
Sponsored
OneDrive
Maps
Show ad
AdChoices
LOS ALTOS / 68°F
NEWS
ENTERTAINMENT
SPORTS
MONEY
LIFESTYLE
HEALTH & FITNESS
FOOD & DRINK
TRAVEL
AUTOS
VIDEO
powered by Microsoft News
WEEK IN AUTOS
100 simple repairs you don't need to go to the shop for
The Family Handyman
Best cars of the 2018 Woodward Dream Cruise
Motor Trend
Car products that are complete wastes of money
Cheapism
John McCain's best Saturday Night Live moments
EW Entertainment Weekly
The location system that connects your world.
Pre-order now
FINDSTER
Closer to you
AdChoices
Stars Remember Neil Simon
The Hollywood Reporter
TRENDING NOW
Jacksonville mass shooting
Sen. McCain has died
McCain to lie in state
Treatment for Mel B.
Frankel: 'Hard to breathe'
Dershowitz on NY probe
New royal dog
Playwright Neil Simon dies
Tiger closes with 70
Hill leaving ESPN
'Game of Thrones' delayed
College football buzz
Jackson Odell cause of death
Weekend box office
Frankel Says It's 'Hard to Breathe' After Shields' Death
US Weekly
Cohen's attorney backs away from claim about Trump
The Washington Post
NFL players react to fatal shooting at 'Madden' tourney
For The Win

Figure 3.1 MSN portal example

chance, given that the design was for an equal split, was extremely unlikely. The investigation discovered that because user engagement increased in the Treatment, some of the most heavily engaged users were classified as bots and removed from analysis. After correcting this bot filtering, the results showed the reverse Treatment effect: user engagement increased by 3.3% in the Treatment!

Bot filtering is a serious problem, especially for search engines. For Bing, over 50% of US traffic is from bots, and that number is higher than 90% in China and Russia.

An SRM check is critical. Even a small imbalance can cause a reversal in the Treatment effect, as the last example shows. SRMs are commonly due to missing users (generally, experiment units) that are either extremely good, such as heavy users, or extremely bad, those users with no click count. This demonstrates that even though the population difference appears small, it can significantly skew the results. A paper on diagnosing SRMs was recently published (Fabijan et al. 2019).

Threats to External Validity

External validity refers to the extent to which the results of a controlled experiment can be generalized along axes such as different populations (e.g., other countries, other websites) and over time (e.g., will the 2% revenue increase continue for a long time or diminish?).

Generalizations across populations are usually questionable; features that work on one site may not work on another, but the solution is usually easy: rerun the experiment. For example, successful experiments in the United States are typically tested in other markets instead of assuming the results will generalize.

Generalizations across time are harder. Sometimes a holdout experiment is left running for months to assess the long-term effects (Hohnhold, O'Brien and Tang 2015). Chapter 19 discusses how to address long-term effects. Two key threats to external validity on a time-basis are *primacy* effects and *novelty* effects.

Primacy Effects

When a change is introduced, users may need time to adopt, as they are *primed* in the old feature, that is, used to the way it works. Machine-learning algorithms may also learn better models and depending on the update cycle, this may take time.

Novelty Effects

Novelty effect, or newness effect, is an un-sustained effect. When you introduce a new feature, especially one that's easily noticed, initially it attracts users to try it. If users don't find the feature useful, repeat usage will be small. A Treatment may appear to perform well at first, but the Treatment effect will quickly decline over time.

An example of something that we are *not* looking for is one told in *Yes!: 50 Scientifically proven ways to be Persuasive* (Goldstein, Martin and Cialdini 2008). In that book, the authors discuss how Colleen Szot authored a television program that shattered a nearly 20-year sales record for a home-shopping channel. Szot changed three words in a standard infomercial line that caused a huge increase in the number of people who purchased her product: instead of the all-too-familiar "Operators are waiting, please call now," it was "If operators are busy, please call again." The authors explain that this is social proof: viewers think "If the phone lines are busy, then other people, like me, who are also watching this infomercial are calling, too."

Ploys, such as the above, have a short shelf life if users recognize that it is used regularly. In a controlled experiment, the analysis will show an effect that quickly diminishes.

Another example is shown in Figure 3.2. The MSN website had a stripe at the top that looked like this (Dmitriev et al. 2017):

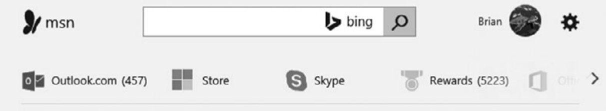

Figure 3.2 MSN page with Outlook.com link

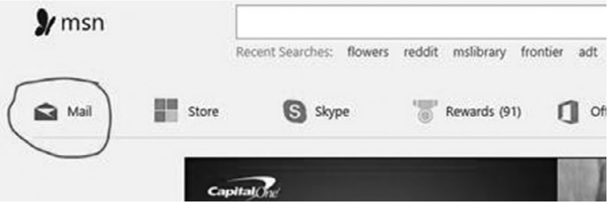

Figure 3.3 MSN page changed to use link to Outlook application

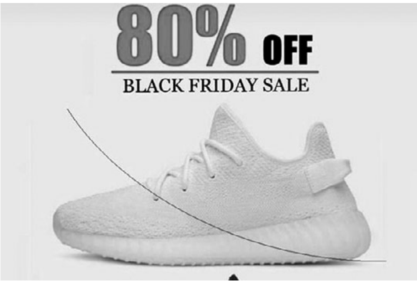

Figure 3.4 Phone ad with fake hair, hoping you'll swipe it off and click-through by mistake

Microsoft changed the Outlook.com link and icon to directly open the Outlook Mail application (Figure 3.3), which gives users a richer, better e-mail experience.

As expected, the experiment showed that more users in Treatment used the Mail app relative to Control, but there was no expectation that the click-through rate would increase. Surprisingly though, there was an extremely large increase of 28% in the number of clicks on that link in Treatment relative to Control. Were users liking the Mail app more and using it more frequently? No. The investigation showed that users were confused that Outlook.com did not open and clicked the link multiple times.

Finally, Chinese sneaker manufacturer Kaiwei Ni had an Instagram ad that showed up on phones with a fake stray hair as shown in Figure 3.4. Users were tricked into swiping on the ad to remove the hair, and many of them clicked through. The novelty effect was likely significant here. More than that, the ad was not only removed from Instagram, but the account was disabled (Tiffany 2017).

Detecting Primacy and Novelty Effects

An important check for primacy and novelty effects is to plot usage over time and see whether it's increasing or decreasing. Take the above MSN example, the percentage of users clicking the Mail link clearly decreased over time, as shown in the graph in Figure 3.5.

The standard analysis of experiments assumes that the Treatment effect is constant over time. This kind of trend is a red flag that indicates a violation of the assumptions. Such experiments need to run longer to determine when the Treatment effect stabilizes. In many cases, and stressed in this example, the insight is enough to declare the idea bad. This approach is simple and effective in most cases, but we must warn you that there are some caveats to watch out for, especially if you do run the experiment a long time (See Chapter 23).

One additional option to highlight possible novelty/primacy effects is to take the users who appeared in the first day or two (as opposed to all users over time) and plot the treatment effect for them over time.

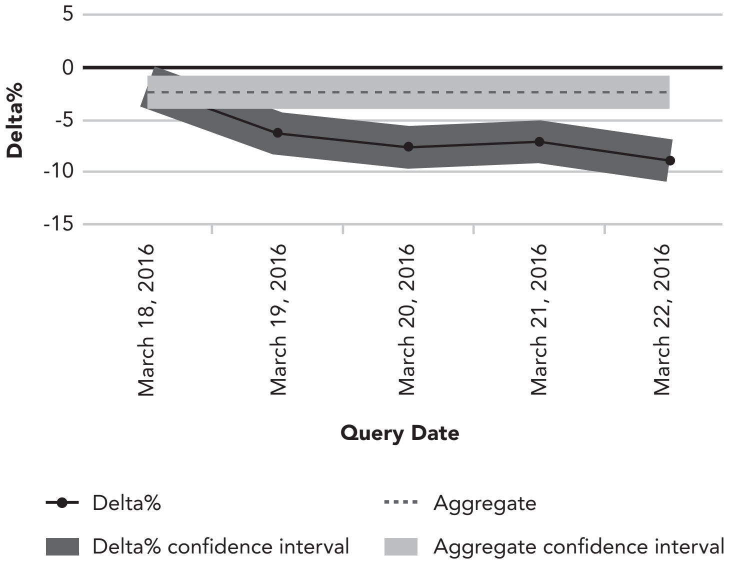

Figure 3.5 MSN user engagement decreasing over time

Segment Differences

Analyzing a metric by different segments can provide interesting insights and lead to discoveries, for which we sometimes invoke Twyman's law and discover a flaw or new insight to help with future iterations of an idea. These are advanced tests with an example that you can address in the later maturity phases of your experimentation system.

What are good segments? Here are several:

- Market or country: some features work better in some countries; sometimes an underperforming feature is the result of poor translation to another language, that is, localization.
- Device or platform: is the user interface on a browser, desktop, or mobile phone? Which mobile platform are they using: iOS or Android? Sometimes the browser version can help identify JavaScript bugs and incompatibilities. On mobile phones, the manufacturers (e.g., Samsung, Motorola) provide add-ons that can cause features to fail.
- Time of day and day of week: plotting effects over time can show interesting patterns. Users on weekends can be different in many characteristics.
- User type: new or existing, where new users are ones that joined after a date (e.g., experiment start, or perhaps a month prior).
- User account characteristics: single or shared account at Netflix, or single vs. family traveler on Airbnb.

Segmented views are commonly used two ways:

1. Segmented view of a metric, independent of any experiment.
2. Segmented view of the Treatment effect for a metric, in the context of an experiment, referred to in Statistics as *heterogeneous* Treatment effects, indicating that the Treatment effect is not homogenous or uniform across different segments.

Segmented View of a Metric

When the click-through rates on Bing mobile ads were segmented by different mobile operating systems, they were very different as shown in the graph in Figure 3.6.

While the initial inclination was to form stories about the loyalty of the users and how the populations differ, an investigation uncovered that this was due to different click tracking methodologies used for different operating systems. There are several ways of tracking clicks, and they differ in fidelity (Kohavi,

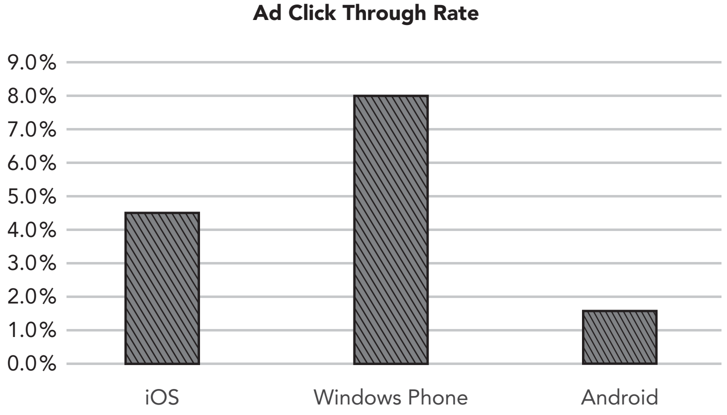

Figure 3.6 CTRs for different mobile Operating Systems

Messner et al. 2010), which leads to different loss rates. On iOS and Windows Phone, a redirect was used to track the clicks, that is, the click always goes to a server, is logged, and is then redirected to the destination. This methodology has high fidelity, but the user experience is slower. On Android, click tracking was done using a web beacon to indicate a click and then redirected the browser to the destination page. This methodology is faster for the user, but lossy; some web beacons will not make it and the link will not be recorded. This can explain the click-through rate (CTR) difference between iOS and Android, but why was the Windows Phone click-through rate so high? The investigation discovered that along with the redirect, there was a bug where user swipes were incorrectly recorded as a click. Bugs happen. When you see anomalous data, think of Twyman's law and investigate the issue.

Segmented View of the Treatment Effect (Heterogeneous Treatment Effect)

In one experiment, a user interface change was made, which resulted in a very strong difference between browser segments. For almost all browser segments, the Treatment effect was a small positive improvement on key metrics, but for the Internet Explorer 7 segment, there was a strongly negative Treatment effect on key metrics. As with any strong effect (positive or negative), you should invoke Twyman's law and drill into the cause. An investigation revealed that the JavaScript used was incompatible with Internet Explorer 7, causing an error that prevented users from clicking links in certain scenarios.

Such insight is only possible when drilldowns into segments are enabled, that is, looking at the Treatment effect for different segments, also referred to in Statistics as Conditional Average Treatment Effects (CATEs). A good overview on Heterogenous Treatment Effects is available at EGAP (2018). Identifying interesting segments, or searching for interactions, can be done using machine learning and statistical techniques, such as Decision Trees (Athey and Imbens 2016) and Random Forests (Wager and Athey 2018).

If you can alert the experimenter to interesting segments, you will find many interesting insights (but remember to correct for multiple hypothesis testing, as noted above). Getting organizations to run A/B tests is an important step; providing them with more information than just the overall Treatment effect gives new insights that help accelerate innovation.

Analysis by Segments Impacted by Treatment Can Mislead

It is possible to evaluate the Treatment effect of two mutually exhaustive and exclusive segments, and see that the OEC increases for both, yet declines overall. Unlike Simpson’s paradox (described in the next section), this is due to migration of users from one segment to another.

For example, assume you have a metric, sessions-per-user, that you care about. You are working on a new product feature F, which few users use, so you focus on users of F and the complement (those not using F). You see that in your experiment, sessions-per-user goes up for users of F. Now you look at the complement and see that their sessions-per-user go up. Can you celebrate? NO! It is possible that sessions-per-user overall decreased or stayed flat.

As an example, users of F average 20 sessions-per-user, while those not using F average 10 sessions-per-user. If the Treatment causes users with 15 sessions-per-user to stop using F, the average sessions-per-user will rise for the segment using F (we removed users with lower than average sessions-per-user), and it will rise for the complement (we added users with higher average sessions-per-user), but the aggregate could move in any direction: up, down, or flat (Dmitriev et al. 2016, section 5.8).

When users move from one segment to another, interpreting metric movements at the segment level may be misleading, so the Treatment effect of the non-segmented metric (aggregate) should be used. Ideally, segmenting should be done only by values that are determined prior to the experiment, so that the Treatment could not cause users to change segments, though in practice restricting segments this way may be hard for some use cases.

Simpson's Paradox

The following is based on Crook et al. (2009). If an experiment goes through ramp-up (see Chapter 15) that is, two or more periods with different percentages assigned to the variants, combining the results can result in directionally incorrect estimates of the Treatment effects, that is, Treatment may be better than Control in the first phase and in the second phase, but worse overall when the two periods are combined. This phenomenon is called Simpson's paradox because it is unintuitive (Simpson 1951, Malinas and Bigelow 2004, Wikipedia contributors, Simpson's paradox 2019, Pearl 2009).

Table 3.1 shows a simple example, where a website has 1 million visitors per day on two days: Friday and Saturday. On Friday, the experiment runs with 1% of traffic assigned to the Treatment. On Saturday that percentage is raised to 50%. Even though the Treatment has a conversion rate that is better on Friday (2.30% vs. 2.02%) and a conversion rate that is better on Saturday (1.2% vs. 1.00%), if the data is simply combined over the two days, it appears that the Treatment is performing worse (1.20% vs. 1.68%).

There is nothing wrong with the above math. It is mathematically possible that $\frac{a}{b} < \frac{A}{B}$ and that $\frac{c}{d} < \frac{C}{D}$ while $\frac{a+c}{b+d} > \frac{A+C}{B+D}$. The reason this seems unintuitive is that we are dealing with weighted averages, and the impact of Saturday, which was a day with an overall worse conversion rate, impacted the average Treatment effect more because it had more Treatment users.

Here are other examples from controlled experiments where Simpson's paradox may arise:

- Users are sampled. Because there is concern about getting a representative sample from all browser types, the sampling is not uniform, and users in some browsers (such as, Opera or Firefox) are sampled at higher rates. It is

Table 3.1 *Conversion Rate for two days. Each day has 1M customers, and the Treatment (T) is better than Control (C) on each day, yet worse overall*

	Friday	Saturday	Total
	C/T split: 99% / 1%	C/T split: 50% / 50%	
C	$\frac{20,000}{990,000} = 2.02\%$	$\frac{5,000}{500,000} = 1.00\%$	$\frac{25,000}{1,490,000} = 1.68\%$
T	$\frac{230}{10,000} = 2.30\%$	$\frac{6,000}{500,000} = 1.20\%$	$\frac{6,230}{510,000} = 1.20\%$

possible that the overall results will show that the Treatment is better, but once the users are segmented into the browser types, the Treatment is worse for all browser types.

- An experiment runs on a website that is implemented in multiple countries, say the United States and Canada. The proportions assigned to the Control and Treatment vary by country (e.g., the United States runs at 1% for the Treatment, while the Canadians do power calculations and determine they need 50% for the Treatment). If the results are combined, the Treatment may seem superior, even though the results were segmented by country, the Treatment will be inferior. This example directly mirrors our previous ramp-up example.
- An experiment is run at 50/50% for Control/Treatment, but an advocate for the most valuable customers (say top 1% in spending) is concerned and convinces the business that this customer segment be kept stable and only 1% participate in the experiment. Similar to the example above, it is possible that the experiment will be positive overall, yet it will be worse for both the most valuable customers and for "less-valuable" customers.
- An upgrade of the website is done for customers in data center DC1 and customer satisfaction improves. A second upgrade is done for customers in data center DC2, and customer satisfaction there also improves. It is possible that the auditors looking at the combined data from the upgrade will see that overall customer satisfaction decreased.

While occurrences of Simpson's paradox are unintuitive, they are not uncommon. We have seen them happen multiple times in real experiments (Xu, Chen and Fernandez et al. 2015, Kohavi and Longbotham 2010). One must be careful when aggregating data collected at different percentages.

Simpson's reversal seems to imply that it is mathematically possible for a drug to increase the probability of recovery in the aggregate population yet decrease the probability (so it is harmful) in every subpopulation, say males and females. This would seem to imply that one should take the drug if gender is unknown yet avoid it if gender is either male or female, which is clearly absurd. Pearl (2009) shows that observational data alone cannot help us resolve this paradox, as the causal model will determine which data to use (the aggregate or the subpopulation). The "Sure-Thing Principal" Theorem (6.1.1) states that if an action increases the probability of an event E in each subpopulation, it must also increase the probability of E in the population as a whole.

Encourage Healthy Skepticism

It had been six months since we started concerted A/B testing efforts at SumAll, and we had come to an uncomfortable conclusion: most of our winning results were not translating into improved user acquisition. If anything, we were going sideways…
– Peter Borden (2014)

Trustworthy experimentation is sometimes tough for organizations to invest in, as it involves investing in the unknown—building tests that would invalidate results if the tests fired. Good data scientists are skeptics: they look at anomalies, they question results, and they invoke Twyman's law when the results look too good.