A Protocol to Reduce Bias & Variance in Head-to-Head Tests

Akim Boyko* Zaid Harchaoui† Thomas Nedelec* Vianney Perchet*‡

Abstract

Head-to-head tests are an effective way to compare the performance of competing approaches. Practical constraints imposed by platforms such as Google AdWords prevent us from using a simple user based split in running these head-to-head tests. Hence, we have developed a new protocol: a geographical split of the total population into two sub-populations allocated to each competitor and time-based swaps of these populations between the two competitors. The proposed protocol is more robust to variability in sales due to seasonal, regional or other latent effects as opposed to simpler splits based only on geography and/or without time-based swaps.

We provide theoretical proof and show empirically that both features are necessary to reduce bias and variance in the estimation of performance of advertising campaigns. We also prove on a simplified model that increasing the number of splits and swaps reduces the bias and variance.

1 Introduction

Comparing the performance of two algorithms, website layouts, recommender systems etc., is a fairly easy and well understood problem when a single entity can control and configure custom tests. In such cases, one can resort to standard A/B test procedures whose guarantees are well studied in the literature [1]. However, the process becomes trickier when one wants to compare two external and non-cooperative solutions. We shall refer to the setting of comparing two external solutions by running a competition between them as a Head-To-Head (H2H) test.

To further concretize, we consider the setting in which an advertiser wants to find an external solution to optimize its online advertising campaign. In order to take an informed decision, the advertiser is planning to organize a H2H to choose the best solution according to some business metrics. During the H2H, the two competitors will be engaged in a competition (over a rather short amount of time, say a couple of months) to determine which one has the better performance. We will denote these two competitors by A and B. At the end of the H2H, the brand will be able to choose the best solution to run its online campaign, say on Google AdWords.

The standard way to set up H2H tests is to expose, at random, half of the population to one or the other solution, for instance based on some "random" hashing of their user id. If the traffic of the website is large enough, the laws of large numbers ensures that both competitors will be allocated roughly the same number of users (the deviation is in the order of the square root of the number of users) and, more importantly, the overall characteristics of the users will be the same in both populations[1]. Using these properties, it is quite straightforward to estimate, without bias and with a

*Criteo Research
†University of Washington
‡Ecole Normale Supérieure Paris-Saclay

[1]Actually, during such a test, the publisher must ensure that "rare events", arrive with a probability of the order of the inverse of the total population. Such events will typically happen in only one of two populations and should be removed to reduce variance.

small variance, the difference in expected performances of the two competitors by simply considering the empirical difference in the performance.

However, sometimes it is not possible to split the population of users at random between the two competitors for several reasons. For example, it is possible that each competitor knows only its own users and there might not be a mapping from them to a global id to ensure a mutually exclusive split. The second more typical constraint is that such H2H tests maybe run on external entities, for example the Google AdWords, that may have some strong limitations which prevent the splitting of the population.

While it is possible to set up H2H by splitting the catalog of products between two competitors, this is hardly desirable as the competitors will most certainly compete for the same users and the performances of both algorithms may suffer drastically from this contention. A possible way to overcome that issue is to create uniform splits of products, independent of the users, such that no products from the two different populations are related to the keywords. Obviously, this is in itself a non-trivial problem, that might not be solvable, as most catalogs contains highly related items.

The most pragmatic split of the population of users is based on the shared features between the two competitors. For the sake of simplicity, we will focus on two such features that are almost always available: time of the request and geography of the user. The first one is self-explanatory. The geographic feature can help discriminate between users at different granularities (zip code, town, regions, states). To avoid existing overlaps between zip codes and towns, we only consider the region or state granularity, depending on the market we are focusing on (the US, Brazil, France, India, etc.). To further concretize this study, we shall assume that the external entity in question is the Google AdWords which has all the constraints listed above.

Given the restriction that users can be allocated to only one competitor based on the time or geographic feature, we would like to construct a family of "admissible" H2H protocols, analyze their empirical and theoretical guarantees, and provide recommendations on the choice of the protocol based on external constraints. We also show that following such protocols is necessary to reduce the high variance and the bias on the performance that can be caused by season and region effects. We emphasize the impact of the seasonal (Figure 1) and regional (Figure 2) variability on the sales revenue by considering the sales revenue earned by Criteo across a wide band of its partners. Given the huge variance, it is imperative to devise a protocol that explicitly minimizes such variance; otherwise, the conclusions that would be drawn from a H2H would be highly conditional on the split used and are not at all indicative of actual performance.

2 A Variance Reducing H2H Protocol

In this section, we describe a family of H2H protocols that can be used to compare the performance between two competitors A and B. As mentioned in the introduction, we assume that users can be partitioned by their geographic area (at the state or region level, depending on the countries) and time (days, weeks or months).

Our protocols are therefore based on the two following partitions:

Geographic Splits: The set of states/regions is partitioned into two non-overlapping subsets $\mathcal{S}_A$ and $\mathcal{S}_B$. Note, that one of the sets can also be empty. Then all users in state in $\mathcal{S}_A$ are allocated to the competitor A and all users coming in state $\mathcal{S}_B$ are allocated to B.

Note that, we do not need any explicit constraints on the geographic split used. The set of states can be unbalanced in terms of population, incomes, etc. We might call a split "fair" if some metrics are aligned on both subsets $\mathcal{S}_A$ and $\mathcal{S}_B$. For instance, a fair split might require

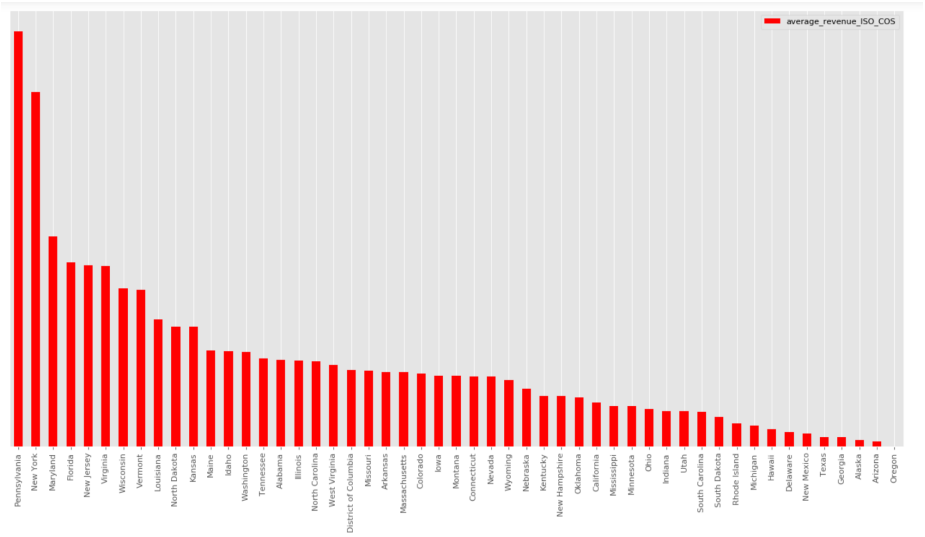

Figure 1: Variability of amount of sales across states for two different partners.

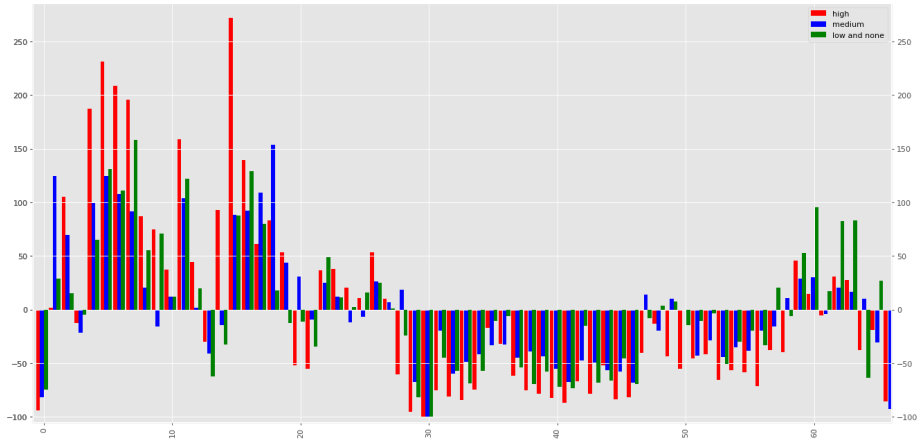

Figure 2: Variability of revenue averaged per week. Products are grouped into three buckets with the same overall total sales based on the price. Products in the red bucket are the most expensive and those in the green bucket least expensive.

that, in the past month, the total number of conversions, money spent, clicks, etc. are similar in both subsets (up to some reasonable deviations). However, our protocol does not require fair splits to be effective.

Time Swaps: Given a geographical split $\mathcal{S}_A, \mathcal{S}_B$ allocated respectively to A and B for a given amount of time, a swap is the opposite allocation ($\mathcal{S}_A$ to B and $\mathcal{S}_B$ to A) for the same amount of time.

Our protocols are defined using N blocks of $2T$ days (obviously, the total length of the protocol is $2NT$ days). During the first block of $2T$ days, we consider the first geographical split $\mathcal{S}_A(1), \mathcal{S}_B(1)$ used during T days, before a swap occurs. During the second block of $2T$ days, the geographical split used is $\mathcal{S}_A(2), \mathcal{S}_B(2)$, again for T days before a swap. This sub-procedure is repeated N times.

As a consequence, our family of protocols is characterized by the two integers N, T and a sequence of splits $\big(\mathcal{S}_A(n), \mathcal{S}_B(n)\big)_{n\in\mathbb{N}}$.

We give an illustration of a H2H protocol below.

	T	$2T$	$3T$	$4T$	...	$(2N-1)T$	$2NT$
Competitor A	$\mathcal{S}_A(1)$	$\mathcal{S}_B(1)$	$\mathcal{S}_A(2)$	$\mathcal{S}_B(2)$	...	$\mathcal{S}_A(N)$	$\mathcal{S}_B(N)$
Competitor B	$\mathcal{S}_B(1)$	$\mathcal{S}_A(1)$	$\mathcal{S}_B(2)$	$\mathcal{S}_A(2)$	...	$\mathcal{S}_B(N)$	$\mathcal{S}_A(N)$

Of course, if often happens that a H2H is organized when a new competitor (say, B) challenges the existing client (say, competitor A) of a publisher. In that case, it is possible to actually get rid of the first swap and to compare the performances of B on $\mathcal{S}_B(1)$ during the first T days with the performances of A also on $\mathcal{S}_B(1)$ but during the T days prior to the head-to-head.

2.1 Experimental Verification

In this section, we provide empirical evidence for the importance of geographic splits and time swaps. We first use the data collected by Criteo and simulate a H2H between a version A of Criteo and a version B of it. Obviously, both algorithms have the same performance (since they are identical). We would like to point out that measuring the differences between the estimated performance of almost identical-competitors is a good proxy to estimate a protocol.

In Figure 3, we illustrate the reduction in variance in measured up-lifts (that should be ideally equal to 0) by using weekly-swaps.

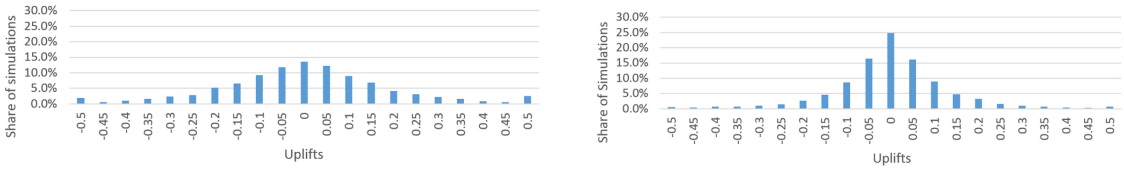

Figure 3: Illustration of possible variance reduction in simulated H2H when using swaps (on the right) vs. no swaps (on the left). Since the same algorithm is used in both populations, higher concentration around 0 is desired.

Figure 2 already illustrated the time variability on a week-to-week basis. We also provide the same picture in Figure 4 where variability is computed on a day-to-day basis. These data provided by a partner of Criteo show that performances can have a strong variability in time. For instance, in our example, campaign performance deteriorated after 20 weeks, sinking below the 1.5 year average performance for 30 weeks and finally ended being above average for the last 10 weeks. Effects are a bit less dramatic when considering daily variability. Figure 2 and 4 give the first intuition on why the frequency of swaps also matters. Assume, for instance, that this partner if allocated to the competitor A during the first half of this 1.5 years period and to competitor B during the second time. Then the former has a strong advantage. With this type of variability, we would typically require at least weekly swaps or even better daily swaps (as the variability effects of "bad" or "good" trends would be reduced).

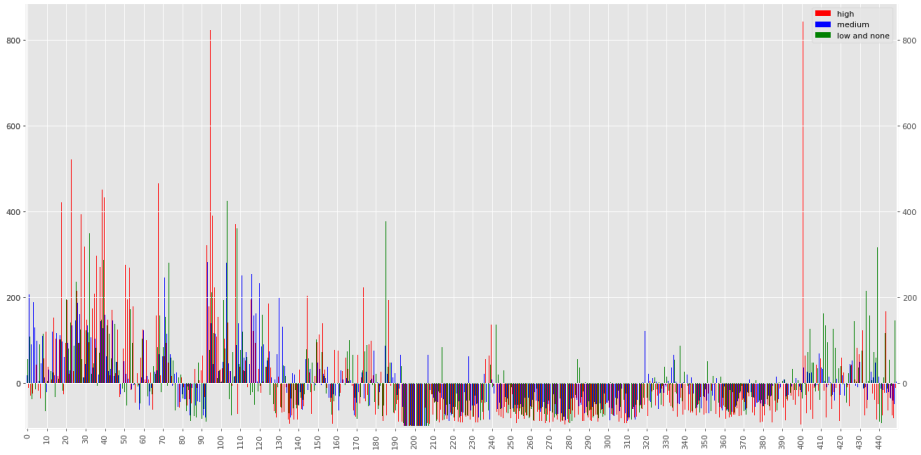

Figure 4: Variability of sales revenue on a day-to-day basis. Products are grouped into three buckets with the same overall total of sales accordingly to their prices: products in the red bucket are the most expensive and those in the green bucket the less expensive

3 Theoretical Verification

We study in this section how geographical splits and time swaps effectively reduce the bias and variance of performances estimators.

We denote by $[K] := \{1, \dots, K\}$ the sets of geographical states and for simplicity[2] we assume that the performances of competitors A and B in state k follows a binomial of parameter $n^{(k)} \in \mathbb{N}$ (the same for both competitors) and $p_A^{(k)} \in [0,1]$ or $p_B^{(k)} \in [0,1]$ respectively for competitors A and B. If state $k \in [K]$ is allocated to competitor A, then we denote its performances by $R_A^{(k)} \in \mathbb{N}$. Notice that the expectation of $R_A^{(k)}$ is equal to $p_A^{(k)} n^{(k)}$

3.1 Geographical Splits Without Time Swaps

Let us assume that the performances are time invariant, i.e., $p_A^{(k)}$ - the probability of conversion - does not evolve through time. As a consequence, the expected performance of competitor A is

$$\mathbf{R}_A = \sum_{k=1}^{K} p_A^{(k)} n^{(k)} = \mathbb{E} \sum_{k=1}^{K} R_A^{(k)}$$

First, we consider a split $\mathcal{S}_A, \mathcal{S}_b \subset [K]$ fixed before-hand, say by competitor A that is "neutral" from its point of view A, that is such that

$$\sum_{k \in \mathcal{S}_A} n^{(k)} = \sum_{k \in \mathcal{S}_B} n^{(k)} \quad \text{and} \quad \sum_{k \in \mathcal{S}_A} p_A^{(k)} n^{(k)} = \sum_{k \in \mathcal{S}_B} p_A^{(k)} n^{(k)}$$

[2]Obviously, in practice, performance might be distributed according to more complicated distributions. However, we are interested in providing intuition and guidelines on how to construct good protocols. So, our simple model is sufficient for that purpose

Given a neutral split to A, then the natural estimator of the performance of A is $\hat{R}_A = 2\sum_{k\in\mathcal{S}_A} R_A^{(k)}$, i.e., twice the empirical performance on its share of states. The multiplicative parameter 2 is the standard important sampling weight, as only half of the total population is observed by A.

Claim 1. Neutral splits are biased

Although it holds that $\mathbb{E}\hat{R}_A = \mathbf{R}_A$, hence this estimator is unbiased for competitor A, the same statement does not hold for competitor B. Indeed

$$\mathbb{E}\hat{R}_B - \mathbf{R}_B = \sum_{k\in\mathcal{S}_B} \mathbb{E}2R_B^{(k)} - \sum_{k\in[K]} \mathbb{E}R_B^{(k)} = \sum_{k\in\mathcal{S}_B} n^{(k)}p_B^{(k)} - \sum_{k\in\mathcal{S}_A} n^{(k)}p_B^{(k)} \ .$$

As a conclusion, a neutral split for competitor A is unbiased only if that split is also neutral for competitor B and there are no valid justification to the existence of such split

Now that we have shown that even neutral splits are biased for one of the competitors, we now consider random splits, where state k is allocated to competitor A with probability $\theta^{(k)}$ (and thus to B with probability $1-\theta^{(k)}$). The natural estimator of the performance of A is then $\hat{R}_A = \sum_{k\in[K]} \frac{\mathbb{1}\{k\in\mathcal{S}_A^{(k)}\}}{\theta^{(k)}} R_A^{(k)}$.

Claim 2. Random splits are unbiased

This is a standard argument of important sampling:

$$\mathbb{E}\hat{R}_A = \sum_{k\in[K]} \mathbb{E}\frac{\mathbb{1}\{k\in\mathcal{S}_A^{(k)}\}}{\theta^{(k)}} R_A^{(k)} = \sum_{k\in[K]} \frac{\theta^{(k)}}{\theta^{(k)}}\mathbb{E}R_A^{(k)} = \sum_{k\in[K]} n^{(k)}p_A^{(k)} = \mathbf{R}_A$$

and the same computations holds for competitor B.

We now compute the variance of the estimator of random splits, with state $k\in[K]$ to competitor A with proba $\theta^{(k)}$

$$\mathrm{Var}(\hat{R}_A) = \sum_{k=1}^{K} \mathrm{Var}\left(\frac{\mathbb{1}\{k\in\mathcal{S}_A\}}{\theta^{(k)}} R_A^{(k)}\right) = \sum_{k=1}^{K} \frac{1}{\theta^{(k)}} n^{(k)}p_A^{(k)}(1-p_A^{(k)}) + \sum_{k=1}^{K} \left(n^{(k)}p_A^{(k)}\right)^2\left(\frac{1}{\theta^{(k)}}-1\right) \ .$$

Notice that the weights $\theta^{(k)}$ that minimize the variance of the difference $\hat{R}_A - \hat{R}_B$ are given by

$$\theta^{(k)} = \frac{\sqrt{p_A^{(k)} + (n^{(k)}-1)\left(p_A^{(k)}\right)^2}}{\sqrt{p_A^{(k)} + (n^{(k)}-1)\left(p_A^{(k)}\right)^2} + \sqrt{p_A^{(k)} + (n^{(k)}-1)\left(p_A^{(k)}\right)^2}} \simeq \frac{p_A^{(k)}}{p_A^{(k)} + p_B^{(k)}}$$

if $n^{(k)}p_A^{(k)} \gg 1$. Thus, if both competitors have approximatively the same state-by-state performances ($p_A^{(k)} \simeq p_B^{(k)}$), the choice of $\theta^{(k)} = 1/2$ makes sense.

3.2 Geographical Splits With Time Swaps

One can now wonder why we need time-based swaps if we already have unbiased estimators. This is to reduce the variance of the final estimators. In the following section, we assume that the H2H

runs on two periods and the expected performances of competitors are the same, state by state, on each period. The natural estimator of performances is the average of the estimators obtained on each period.

We denote by $R_A^{(k)}(t)$ the empirical performance of competitor A in state $k \in [K]$ during period t. Its expectation is $n^{(k)}p_A^{(k)}(t) = n^{(k)}p_A^{(k)}$ using a stationarity assumption on the performances.

Claim 3. Swaps remove the bias of fixed splits

Given a fixed split $\mathcal{S}_A, \mathcal{S}_B$, we assume that states in $\mathcal{S}_A$ are allocated to A during the first period and to B during the second one. In that setting, the estimator of $\mathbf{R}_A$ is then

$$\hat{R}_A = \sum_{k \in \mathcal{S}_A} R_A^{(k)}(1) + \sum_{k \in \mathcal{S}_B} R_A^{(k)}(2)$$

which is obviously unbiased since

$$\mathbb{E}\hat{R}_A = \sum_{k \in \mathcal{S}_A} \mathbb{E}R_A^{(k)}(1) + \sum_{k \in \mathcal{S}_B} \mathbb{E}R_A^{(k)}(2) = \sum_{k \in \mathcal{S}_A} n^{(k)}p_A^{(k)} + \sum_{k \in \mathcal{S}_B} n^{(k)}p_A^{(k)} = \mathbf{R}_A$$

Claim 4. Swaps reduce the variance of random splits

Consider two randoms splits performed sequentially with the same probability of allocation $\theta^{(k)}$. Then the variance of the average estimator is:

$$\mathrm{Var}\Big(\frac{\hat{R}_A(1) + \hat{R}_A(2)}{2}\Big) = \sum_{k=1}^{K} \frac{1}{2\theta^{(k)}} n^{(k)}p_A^{(k)}(1 - p_A^{(k)}) + \frac{1}{2}\sum_{k=1}^{K} \big(n^{(k)}p_A^{(k)}\big)^2 \Big(\frac{1}{\theta^{(k)}} - 1\Big)$$

On the other hand, if the first random geographic split is followed by a time swap, then the variance of the estimator is:

$$\begin{aligned}\mathrm{Var}\Big(\frac{\hat{R}_A(1) + \hat{R}_A(2)}{2}\Big) = \mathrm{Var}\Big(\frac{\sum_{k \in \mathcal{S}_A} 2R_A^{(k)}(1) + \sum_{k \in \mathcal{S}_B} 2R_A^{(k)}(2)}{2}\Big) &= \mathrm{Var}\Big(\sum_{k \in [K]} 2R_A^{(k)}(1)\Big) \\ &= \sum_{k=1}^{K} n^{(k)}p_A^{(k)}\big(1 - p_A^{(k)}\big)\ .\end{aligned}$$

We emphasize that the variance of a geographic split followed by a time swap (this on two periods) is exactly equal to the variance obtained when estimating on the full population during one period.

In particular, if $\theta^{(k)} = \frac{1}{2}$, then the variance reduction is in the order of $\frac{1}{2}\sum_{k=1}^{K}\big(n^{(k)}p_A^{(k)}\big)^2$ that can actually be of the same order of $\mathbf{R}_A^2$. This would happen if one state say k^* (or a few of them) has an overwhelming importance, i.e., if $n^{(k^*)}p_A^{(k^*)} \gg n^{(k)}p_A^{(k)}$. In that case, the standard deviation of the estimator is of the same size of estimator itself, hence the estimation is meaningless. This non-uniformity across states is actually quite frequent. This phenomenon was already illustrated by Figure 1. We provide in Figure 5 a similar plot using the data of another partner of Criteo, to indicate that time-based variability is indeed quite common.

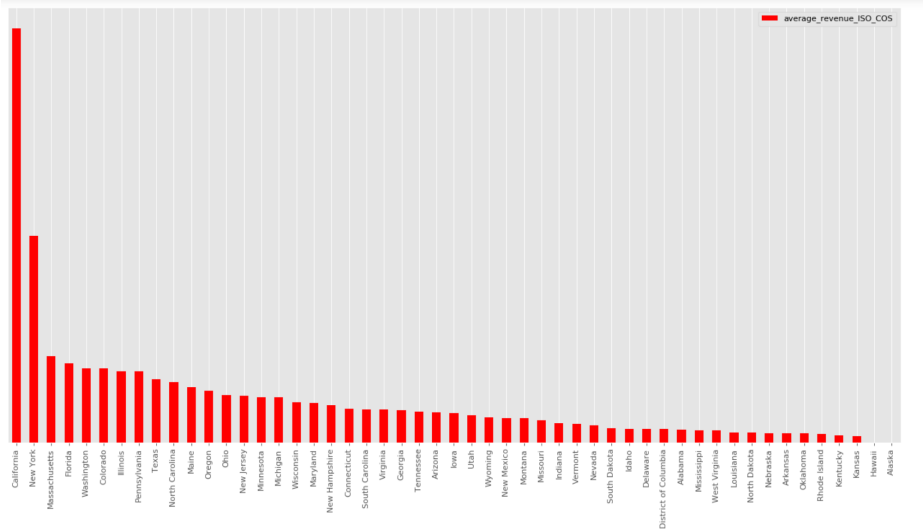

Figure 5: Variability of amount of sales across states for two different partners.

One might argue that by choosing $\theta^{(k)}$ close to 1, the variance of $\hat{R}_A$ without swap would be smaller than the variance with a swap by a factor of roughly $1/2$. This is obviously correct but, on the other hand, the variance of $\hat{R}_B$, for the other competitor, would explode to infinity.

If we no longer assume that the performances of both competitors are stationary in time, then using a fixed split $(\mathcal{S}_1, \mathcal{S}_2)$ where $\mathcal{S}_1$ is allocated to competitor A during the first period and to competitor B during the second one (and the contrary for $\mathcal{S}_2$) might be biased. The simple answer to that question is simply to choose the first allocation *at random* with a coin flip. With probability 0.5, $\mathcal{S}_1$ is allocated to first A then B and with probability 0.5, allocated first to B then A.

3.3 On the frequency of swaps

In this section, we focus on the effect of having several swaps. As we proved that the split is irrelevant, as soon as swaps are involved, we may as well assume that all states are allocated altogether either to competitor A or B; this corresponds to the case where $\mathcal{S}_A = [K]$ and $\mathcal{S}_B = -\emptyset$. And to simplify notation, we may even assume that there is only one state, so that the exponent k can be dropped.

We are now going to assume that the performance of competitor A and B on the unique state evolves with time. During the period t, we denote by $p_A(t)$ and $p_B(t)$ the probability of conversion of competitor A and B. We assume on the other hand that the population n remains constant. In the following, the noise would only reinforce our claims, so assume that we are in a noiseless scenario.

Claim 5. The frequency of swaps must be adapted to the seasonality effects

We assume that swaps can only happen every day, so the unit of time is going to be a day. Note that the results hold regardless of the time unit used.

Example 1. Assume that the performance increase linearly in time, i.e., $p_A(t) = p_A + \varepsilon t$ and $p_A(t) = p_B + \varepsilon t$ for period $t = 0, 1, \ldots$ This implies that the performance of competitor A at time t is equal to $np_A(1 + \varepsilon t)$ if the state is allocated to it and is 0 otherwise.

Assume that there is a swap every τ days, i.e., competitor A controls the market during days $t = 0$ to $t = \tau - 1$, then from $t = 2\tau$ to $t = 3\tau - 1$, etc. and that the H2H is run during $2T\tau$ days.

Then the average performance of A is

$$\hat{R}_A = \frac{1}{T\tau} \sum_{s=0}^{T-1} \sum_{t=0}^{\tau-1} p_A + \varepsilon(2s\tau + t) = p_A + \varepsilon(T\tau - \frac{\tau}{2} - \frac{1}{2})$$

On the other hand, the average performance of B is

$$\hat{R}_B = \frac{1}{T\tau} \sum_{s=0}^{T-1} \sum_{t=0}^{\tau-1} p_B + \varepsilon((2s+1)\tau + t) = p_B + \varepsilon(T\tau - \frac{\tau}{2} - \frac{1}{2}) + \varepsilon\tau$$

Note that the real average performance of A is actually

$$\mathbf{R}_A = \frac{1}{2T\tau} \sum_{t=0}^{2T\tau-1} p_A + \varepsilon t = p_A + \varepsilon(2T\tau - 1)$$

We can arrive at two conclusions from this example:

- The bias in the estimation decreases with the frequency of swaps ($1/\tau$).
- If $p_A > p_B$, the competitor A outperforms competitor B. But if $\varepsilon\tau > p_A - p_B$ then the H2H would indicate the contrary. This implies that daily swaps ($\tau = 1$) can output the right conclusion of the H2H while weekly swaps ($\tau = t$) might overestimate the performance of B that might win the H2H. Performing a weekly swap then is better than not swapping at all.

Example 2. Assume that the performances are constant $p_A(t) = p_A$ except for a short amount of time, between $t = \tau$ and $t = \tau + \delta$ where it explodes $p_A(t) = p_A^* \gg p_A$. One might think of sales, vacations, etc. To be more practical, assume that $\delta = 7$, i.e., sales last for 1 week.

If weekly swaps are used, it might be the case that the rare event happens on a week allocated to B only. Thus its average performance over $2T$ weeks is

$$\hat{R}_B = p_B + \frac{p_B^* - p_B}{T} \quad \text{while} \quad \hat{R}_A = P_A \ .$$

So if $p_A > p_B$ but p_B^* is big enough, then the head-to-head might output B as a winner.

On the other hand, if daily swaps are used, then

$$\hat{R}_B = p_B + \frac{4(p_B^* - p_B)}{7T} \quad \text{while} \quad \hat{R}_A = P_A + \frac{3(p_A^* - p_A)}{7T} \ .$$

It is still possible that B, by pure luck, is declared winner of the head-to-head, but the increase of performance $p_B^* - p_B$ must roughly be 7 times bigger than with weekly swaps for this to happen.

4 Our H2H Protocol Recommendation

Our main message with all of the above analysis is the following. One time-based swap is absolutely necessary but the number of geographical splits is less important. The H2H can be composed of one, two or more splits, the procedures will still work and, as mentioned before, the more the better.

If there exists already historical data on competitor A, then it is possible to perform the H2H by assigning only once a set of states to competitor B. Its performance on this set of states will then be compared to the past performances of A, obviously on the same set of states. This basically reduces to our protocol with one swap restricted to a subset of states. Of course, this protocol is probably the worst one as the bias and variance can be huge (yet it is still better than finding a "neutral" split without swap), but we mention it to illustrate the effectiveness of our protocol family.

Properties of different protocols are summarized in the table below

	Fixed split w.o. swap	Random split w.o. swap	Weekly swaps	Daily swaps	Product split w./w.o. swap	User split w.o. swap
Bias:	Huge	None	None	None	Huge	None
Variance:	Small	Huge	Moderate	Small	Small	Small
Effects:	None	Unfair splits	Sales week	Attribution Exposure	Conflict	Implement

Indeed:

- With random split without swaps, it might happen that a competitor is allocated to a vast majority of states while the other gets only a few.
- Weekly swaps might not be frequent enough to handle special weeks of sales, especially if there is only one of them. In that case, the competitor with the "best" states has a strong advantage.
- With daily swaps, it is possible that a user is exposed to both competitors. Then one of them might leverage the performances of the other and/or is attributed a conversion due to the other.
- Product splits generate conflict between competitors that will probably display two similar products from the same partner and second price themselves, degrading their performances.
- User splits would be ideal but cannot be implemented at the moment.

Based on the above, the protocol we recommend is the following

- Based on historical data, find one (or more) split that look neutral (at least for one of the competitors). If this is impossible, consider one (or several) random split(s).
- Figure out the duration of the seasonality effects and/or if there exists some rare events (such as sales period) during the H2H. Based on these indicators, decide the length of the time-swap. Typically, the more frequent the swap the better the reduction in bias and variance. On the other hand, too frequent swaps might alter the process of allocating the conversions to the right competitor on the client side. Based on our experimental validation, we recommend daily swaps or weekly swaps.
- Once the geographic splits and the time-length of swaps are fixed, use the first geographic split, then swap it. If there is another possible split, follow the same procedure with that one, or re-use the first split, etc.

5 Conclusion

We have introduced a family of estimation protocols based on two widely available features: geographic location of user and time of ad request. We have proved that swapping is necessary to get unbiased estimates and to drastically reduce the variance of the random splits estimate, especially in the presence of seasonality effects or rare events. Such, seasonal and regional effects are typical in ad campaign performance data and hence using a protocol that ignores these effects will typically result in non-conclusive tests.

References

[1] E. Kaufmann, O. Cappé and A. Garivier. *On the Complexity of A/B Testing*, JMLR: Workshop and Conference Proceedings. V.35, 1–23, 2014.