Diluted Treatment Effect Estimation for Trigger Analysis in Online Controlled Experiments

Alex Deng
Microsoft
One Microsoft Way
Redmond, WA 98052
alexdeng@microsoft.com

Victor Hu
Microsoft
One Microsoft Way
Redmond, WA 98052
vihu@microsoft.com

ABSTRACT

Online controlled experiments, also called A/B testing, is playing a central role in many data-driven web-facing companies. It is well known and intuitively obvious to many practitioners that when testing a feature with low coverage, analyzing all data collected without zooming into the part that could be affected by the treatment often leads to under-powered hypothesis testing. A common practice is to use triggered analysis. To estimate the overall treatment effect, certain dilution formula is then applied to translate the estimated effect in triggered analysis back to the original all up population. In this paper, we discuss two different types of trigger analyses. We derive correct dilution formulas and show for a set of widely used metrics, namely ratio metrics, correctly deriving and applying those dilution formulas are not trivial. We observe many practitioners in this industry are often applying approximate formulas or even wrong formulas when doing effect dilution calculation. To deal with that, instead of estimating trigger treatment effect followed by effect translation using dilution formula, we aim at combining these two steps into one streamlined analysis, producing more accurate estimation of overall treatment effect together with even higher statistical power than a triggered analysis. The approach we propose in this paper is intuitive, easy to apply and general enough for all types of triggered analyses and all types of metrics.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Experiment Design

General Terms

Measurement; Experimentation; Design; Theory

Keywords

Controlled experiment; A/B testing; Feature Coverage; Variance reduction; Dilution; Metric

WSDM'15, February 2–6, 2015, Shanghai, China.

http://dx.doi.org/10.1145/2684822.2685307 .

1. INTRODUCTION

In the recent decade, online controlled experiment, also known as A/B testing, has become a must for most of data driven web facing companies. The strength of A/B testing lies in the ability to establish causal relationship between the treatment tested and the effect using simple data collection mechanism that can be implemented as embedded components into existing web services (2). Unlike other statistical methods for causal inference which typically rely on strong assumptions that cannot be verified and therefore tend to produce more false assertions than expected (4), A/B testing makes few assumptions. It is also intuitively easy for non-statisticians to understand. In A/B testing, experimentation units, usually users, are randomly split into control group and treatment group. Traffic in the two groups are then exposed to two different versions of the web service. After collecting data from the two groups over a period of time, we analyze the data at the end of the experiment to compare the two groups, see Figure 1[1]. Because of the random traffic splitting, the two groups are the same by design except for exposure to different versions of the web service. Any true difference between the two must come from the difference of the two versions. In the language of Rubin Causal Model, randomization makes sure that the treatment assignment is *ignorable* (5).

In online controlled experiment, we typically use standard statistical inference methods. There are often two goals:

1. Hypothesis testing: Test the hypothesis that there is no difference between the treatment and the control. If the hypothesis is rejected, we know the treatment has an effect. The confidence level is typically controlled at 5% false positive(claiming an effect by mistake).
2. Point estimation: Estimate the treatment effect. The result is usually presented in a form of 95% confidence interval.

These two goals are deeply connected. In fact knowing confidence interval gives an alternative hypothesis testing method by checking whether 0 is in the 95% confidence interval. In this connection, point estimation goes beyond the simple Boolean claim on whether there is a treatment effect by also providing a possible range for the treatment effect and is therefore more preferable. However, in some cases pure hypothesis testing is much easier to conduct than providing confidence intervals. When testing a feature with low coverage, hypothesis testing of the treatment effect can

[1]Although the name A/B testing might suggests there is only one treatment and one control, it can be extend to multiple treatments and that is very common in practice.

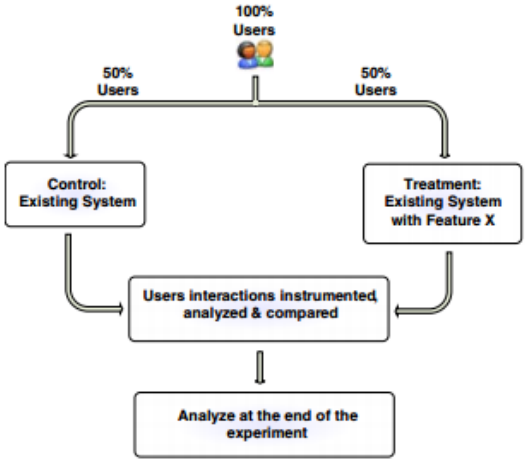

Figure 1: Illustration of Online A/B testing as in (2). Experimentation unit here is user but can be changed to other units such as page-views or visits.

be more efficiently conducted using trigger analysis focusing only on "subset of the data that could be impacted by the treatment"(we will give definition to what this means in Section 2). This is intuitively obvious because by focusing on the triggered subset, the treatment effect is more concentrated and hence stronger. Put it in the other direction, keeping a subset of the data that couldn't be impacted by the treatment is like adding extra noise and will dilute the treatment effect. Analyzing on triggered population is called trigger analysis, in contrast with all up analysis on the original all up or overall population. Hypothesis testing in trigger analysis is also a testing procedure for all up analysis. This is because if there is a treatment effect for triggered subset, there must be a treatment effect for the overall population. Trigger analysis also can provide point estimation for the treatment effect on the triggered population. To provide a point estimation for the overall treatment effect on the all up population, we need to translate our estimate of triggered effect back to the overall effect. This translation is also called dilution because the overall treatment effect is the triggered treatment effect diluted by untriggered complement. Because of this, we also call this overall treatment effect estimation after translation *diluted treatment effect*. The translation step seems straightforward to many people at first glance, and indeed it is for some metrics. But as we will show in Section 2 and Section 3, it turns out to be difficult for a large set of common metrics such as session success rate per user(SSR). Session success rate per user is defined as the average over all users' observed session success rate. Note that at each experiment unit level (user in this case), session success rate is a ratio of the number of successful sessions to the number of sessions, hence the name ratio metrics. These kind of metrics are widely used in search engine evaluation. They are also called double averaging metrics because each user's session success rate is by itself an average. It is different from session success rate (per session) which is defined as the ratio of total successful sessions to the total number of sessions. The double averaged ratio metrics has the advantage of being more robust against outliers as users are treated equally instead of allowing a few super heavy users dominate the metrics with high weightings. Ratio metrics are very common in online A/B testing especially when giving each experiment unit equal weight is preferred. The goal of this paper is to show how to do point estimation for overall treatment effect for all metrics including ratio metrics. We will get it right and get it in an elegant way.

Traditionally majority of interest has been focused on the hypothesis testing side. This is because for many applications of A/B testing, people just need to tell whether or not there is a treatment effect and more importantly, if there is any, whether treatment is better or worse. We call this type of A/B testing *action centric A/B testing*. In recent years, experienced by both authors working in Microsoft, there is a growing demand in what we called *value centric A/B testing*, in which we are also interested in estimating the overall treatment effect of a feature. There are several reasons behind this trend. First is for ROI(return of investment) calibration. Knowing the value of a feature and cost of a feature allows us to know the ROI. If we need to make choices between two or more features with different cost, ROI could help us make final decision. Also in a lot of cases not all treatments with positive effect are worth shipping. New features almost always come with maintenance cost and we should only ship features that are *practically significant*, not just statistically significant. Second reason is for team commitment and performance evaluation. In a data driven culture, part of team performance should be evaluated using objective measurement such as total contribution of features shipped measured by some metrics. A team would often commit to improve certain metric, their OEC(overall evaluation criteria) by x%. This is a crucial step and in some sense a manifesto of a truly data driven culture. For this to be possible, point estimation of overall treatment effect needs to be reported and recorded for each feature shipped. Triggered treatment effect is of little use here because it is often much easier to achieve a large movement (3), and we need to measure impact on the overall population for fair comparison across different teams.

The paper is organized as follows. Section 2 introduces two common types of trigger analyses used in search engine evaluation. We also show why translation from triggered effect to diluted effect is only straightforward for additive metrics like sessions per user and queries per user, and could be involved for ratio metrics such as Session Success Rate per user. We derive exact dilution formula in Section 3 with mild assumptions. The heart of this paper is Section 4 where we introduce a unified approach for all treatment effect dilution problems from a different angle by realizing the whole purpose of trigger analysis with effect dilution is nothing but trying to provide an unbiased estimator for the overall treatment effect with smaller variance. Therefore, instead of tackling the effect dilution problem, we directly tackle the problem of estimating overall treatment effect with extra information of feature triggering. This new approach always provides unbiased point estimation and confidence interval for the overall treatment effect. Moreover, it also quantifies the benefit of using triggering information from a variance reduction perspective. Section 5 provides empirical results from applying this approach on real experimentation data, followed by conclusion in Section 6. In addition, we put a

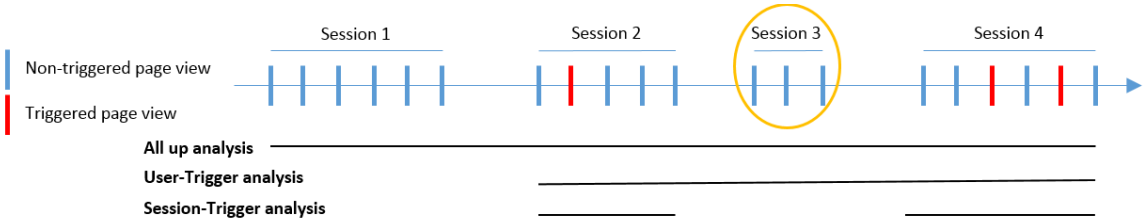

Figure 2: Different types of triggering analysis. User-trigger analysis will analyze from session 2 to session 4. Session-trigger analysis will remove session 1 and session 3. The yellow circled session 3 is a session that is included in user-trigger analysis but not in session-trigger analysis

toy example in Appendix to illustrate the steps of different methods.

Our contribution in this paper includes:

1. We show common mistakes of applying dilution formula to ratio metrics.

2. We derive the exact dilution formula for both additive metrics and ratio metrics. The derivation uses a novel method that we believe is worthy of sharing.

3. The unified overall treatment effect estimation approach is a fresh look of the dilution problem. It applies to all metrics. Comparing to dilution formula, it is easier to implement, requires less assumptions and also less error prone.

2. TRIGGER ANALYSIS

Many online services, especially a complex technology such as search, has many features that only affects a small proportion of traffic. Feature coverage is the proportion of traffics that trigger the feature. For example, a better search result for weather prediction only triggers when user search something related to weather. Some low coverage features trigger for less than 0.1% traffics.

For a low coverage feature, straightforwardly perform hypothesis testing on all up data can suffer from low statistical power. Unless the treatment effect on those triggered traffic is very large, overall effect after diluted by those traffic without feature triggering could be extremely small. In search engine feature experiment in Bing, overall effect for low coverage feature is typically smaller than 0.1%. Also see Rule #2 in (3). Even if all up analysis does show statistically significant result, the accuracy of this estimation could be much lower than the triggered treatment effect. For a feature with coverage less than 20%, we require team to also provide trigger analysis.

Definition of triggering at each impression level (e.g.page-view or query-view) is conceptually easy. We know exactly whether a feature triggered or not in treatment, and with proper counterfactual logging, we should be able to tell whether a feature *would be triggered if it were in treatment* for control group. However, many metrics are only defined at experiment unit level, typically user in online A/B testing. To define many metrics properly, we cannot simply filter out triggered impressions for trigger analysis. To see that, a session is defined as a set of consecutive impressions from a user and we define a successful session if we believe user completed his or her "task" successfully in the session[2]. And if we filter down to triggered impressions, we might only get a subset of impressions for each sessions of a user. Without the holistic view of the whole session we cannot properly define session success. Another example is that sometimes a feature is triggered to help user complete their task easier in follow on impressions within the same session. For instance, speller correction suggestion is triggered when we detect a potential misspelling. A user can click the suggestion and get better results if that is what they really want to query and achieve task completion on that next corrected query(see Figure 3). Although we draw heuristics from search evaluation, we believe similar arguments apply to other domain of online A/B testing.[3]

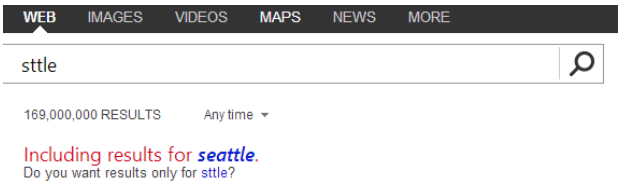

Figure 3: Speller correction for query "sttle".

Due to the reason aforementioned, in search engine evaluation and many other web services, we need to keep session intact when doing trigger analysis. This leads to the concept of session-trigger analysis, in which we take the subset of sessions with at least one triggering event. In particular, users who never trigger the feature will be excluded from the analysis. See Figure 2 for an illustration. However, some feature might have long lasting effects that will not only affect the current session but also affect subsequent sessions. In this case, similar to the logic that we want to maintain the whole session structure when doing trigger analysis, we might also want to keep all sessions after the first triggering event of a user. We can remove sessions before the first triggering event for obvious reason that a feature cannot impact anything before the first exposure. This leads us to the second type of trigger analysis. We call it user-trigger analysis. In Figure 2, we illustrate the difference between all up analysis, session-trigger and user-trigger analysis. The first trigger event is in session 2, so session 1 will be excluded from the analysis. Session 3 is included in user-trigger analysis but not in the session-trigger analysis.

[2]Detailed definition of session and success is out of the scope of this paper.

[3]In other web services, a session normally represent a visit.

SS	Number of Sessions
SatSS	Number of Satisfied/Success Sessions
Tr	Subscript or prefix for triggered
UnTr	Subscript or prefix for untriggered
Overall	Subscript for all up analysis and effect
N	Number of experiment units, typically users
Denom	Denominator

Table 1: Common notations used in formula derivation and variance reduction models.

In practice, user-trigger analysis is more widely used because it applies to general scenarios without assuming the treatment effect only last within the same session the treatment triggered. Using user trigger analysis also make other user based metrics such as sessions per user and queries per user available. If users are very satisfied with a new feature, he or she would return more often to perform more tasks. Those treatment effect impacted sessions might or might not trigger the feature again. In session-trigger analysis, the metric sessions per user merely counts sessions that trigger the feature and therefore loses its meaning of measuring general user engagement. The advantage of using session-trigger is more sensitive by further zooming into the feature impacted subset. Session-trigger analysis supports all session based metrics such as session success rate per user, session time to success per user, as well as all page based metrics such as click through rate, conversion rate, and other ratio metrics. In Section 5 we show for low coverage features, session-trigger analysis can outperform user-trigger analysis. Of course it is important to keep in mind that for session-trigger analysis we assume the treatment has no impact on sessions without triggering. This is also an assumption that we can verify by doing trigger-complement analysis, in which we analyze the complement data of session-trigger analysis and test whether there is indeed no evidence of any impact.

2.1 Effect Dilution

In the context of trigger analysis, assume we have an estimate of the treatment effect in terms of a delta $\Delta_{Tr}(X)$, calculated as the difference of a metric X between treatment group and control group. Here the subscript "Tr" stands for "trigger" and we will use "Tr" and "UnTr" as shorthand for "trigger" and "untrigger". We will use these shorthand that should make sense with contexts. See Table 1 for a complete reference. If Δ_{Tr} is an unbiased estimator for the treatment effect on triggered population(triggered user or triggered sessions), how can we translate it into an unbiased estimator for the treatment effect on the overall population?

A formula that is frequently used for user-trigger analysis is:

$$\Delta_{overall} = \Delta_{Tr} \times \frac{N_{Tr}}{N}, \tag{1}$$

where N_{Tr} and N are user counts in the trigger analysis and all up analysis respectively and for session-trigger N_{Tr}/N is an estimate of the user-trigger rate or feature user coverage. A similar formula for ratio metrics in a session-trigger analysis takes a similar form:

$$\Delta_{overall} = \Delta_{Tr} \times \frac{N_{Tr}}{N} \times \overline{\left(\frac{\text{TrDenominator}}{\text{Denominator}}\right)} \tag{2}$$

where "Denominator" is the denominator of the ratio metrics such as sessions or pages. For session success rate per user, the additional multiplier $\overline{(\frac{TrSS}{SS})}$ is the average session trigger rate for triggered users. For click through rate per user, that would be the average page trigger rate for triggered users where the triggering is defined at the whole session level, i.e. all page views in a triggered session counted as triggered.

There are simple heuristics behind formula (1). For user-trigger analysis, if the treatment effect estimated on the triggered population is Δ_{Tr} and this population is just N_{Tr}/N of the total population, and if the treatment effect on untriggered population is 0, then the overall treatment effect has to be the triggered treatment effect Δ_{Tr} diluted by the trigger rate. For ratio metric in a session-trigger analysis, the same user dilution first need to be applied. Then within triggered users, we still need to dilute by the fact not all sessions are triggered. An intuitive choice seems to be using the average trigger rate of the ratio metric's denominator.

A closer look at this argument above begs more explanation, especially for the session-trigger formula (2). In fact these two formulas are both problematic in general for ratio metrics like session success rate per user, as exemplified by the following using session time to success per user (TTS). Time to success is measuring how fast users get to a satisfied result.[4]:

1. User 1: 1 session-triggered session saving 100(msec), 1 non-session-triggered session, so no savings. User level average TTS saving is -50(msec).
2. User 2: 1 session-triggered session saving 20(msec), 9 non-sesssion-triggered sessions, so no savings. User level average TTS saving is therefore -2(msec)

Taking average of the two users, $\Delta_{Overall} = -26$(msec). However, using formula (2), we first compute $\Delta_{Tr} = -(100+20)/2 = -60$(msec), and then diluted it into $-60 \times 1 \times (1/2+1/10)/2 = -18$(msec). We get very different estimates. In fact, we observed some cases that when we applied diluted formula and compare the result $\Delta_{Overall}$ to the 95% confidence interval we got straight from all up analysis, the formula estimate did not even fall into the interval!

This example exposed the intuition why the formula could fail. When we reason about diluted the average triggered effect by $\overline{(\frac{\text{TrDenominator}}{\text{Denominator}})}$, we are implicitly making a lot of assumptions. Roughly speaking because the treatment effect is itself an average, the dilution by the triggered rate of the denominator should be applied to each individual user level treatment effect first before the average is taken. This could be different from the formula, if there is some correlation between the treatment effect and the trigger rate of the denominator; See Section 3.3.

In the next section, we will derive a rigorous universal formula by using potential outcome framework(5). This framework will enable us to see through the problem in a crystal clear lens. We will suggest how to implement the formula. In Section 4 we argue that variance reduction framework provides a much general implementation that will supersede the formula. Nevertheless the discussion in the next section paves the way for a deeper understanding.

[4]The original example was contributed by our colleague Aron Inger and Ron Kohavi and we made some minor modification.

3. EXACT DILUTION FORMULA

3.1 Rubin's potential outcome framework

Rubin potential outcome framework, also known as Rubin causal model(RCM)(5) is an approach to analyze causal effect based on a conceptual potential outcome pair. For each subject, say user, if there is a causal effect caused by certain treatment, we model this by a pair (X_T, X_C). X_T and X_C is the measurement if the subject is given treatment or not, respectively. And the difference $X_T - X_C$ is the causal effect of the treatment. Note that both X_T and X_C could be nondeterministic and $X_T - X_C$ could be random. The goal of the causal inference is therefore to estimate the *average treatment effect*(ATE) over a population. ATE is

$$\mu = \mathbb{E}(X_T - X_C).$$

In reality, there is no way we can observe the pair. If a subject is exposed to treatment, we observe X_T and not X_C, and vice versa. The missing part is called *counterfactual*.

In online A/B testing, we can estimate $\mathbb{E}X_T$ and $\mathbb{E}X_C$ seperately using sample average in each group. This is because both groups are sampled from the same population. Using $\mathbb{E}(X_T - X_C) = \mathbb{E}X_T - \mathbb{E}X_C$, we see that $\Delta = \overline{X_T} - \overline{X_C}$ is an unbiased estimator for the ATE.

The procedure above is an example of the following rule of thumb:

> To estimate causal effect, we can always transform an unbiased estimator based on potential outcome pairs into an unbiased estimator using observations from treatment group and control group in a randomized experiment.

The advantage of deriving a formula under the potential outcome framework is that we reduce the problem of having two groups of subjects into only one group. This simplification makes the derivation a lot easier and lucid.

3.2 Additive Metrics

Metrics like Queries per user or sessions per user satisfy the following additivity property. For any user, if X is the metric in all up analysis, TrX is that in triggered analysis and UnTrX for trigger-complement analysis, then $X = TrX + UnTrX$. Metrics like session success rate per user or click through rate per user does not have this property since the numerator and denominator are both additive but the ratio is not.

This property is very crucial for a simple formula like (1) to work. To see that, under the potential outcome framework, each user has a potential pair $(X_{iT}, X_{iC}), i = 1, \ldots, N$ and the estimated treatment effect is defined as $\Delta_{overall}(X) = \sum(X_{iT} - X_{iC})/N$.

$$\begin{aligned}
&N \times \Delta_{overall}(X) = \sum(X_{iT} - X_{iC}) \\
&= \sum_{UTr}(X_{iT} - X_{iC}) + \sum_{Tr}(X_{iT} - X_{iC}) \\
&= \sum_{UTr}(X_{iT} - X_{iC}) + \\
&\sum_{Tr}(TrX_{iT} - TrX_{iC}) + \sum_{Tr}(UnTrX_{iT} - UnTrX_{iC}) \\
&= \sum_{Tr}(TrX_{iT} - TrX_{iC}),
\end{aligned}$$

where Tr and $UnTr$ in the subscript of the sum indicate triggered users and untriggered users. The third equation uses the additivity and the last equation uses the fact that there is no treatment effect for untriggered users and even for triggered users, there is no effect on trigger complement.

Divide both side by N and we get the formula (1)

$$\begin{aligned}
\Delta_{overall}(X) &= \frac{\sum_{Tr}(TrX_{iT} - TrX_{iC})}{N_{Tr}} \times \frac{N_{Tr}}{N} \\
&= \Delta_{Tr}(X) \times \frac{N_{Tr}}{N}.
\end{aligned}$$

We make the remark that this formula is correct for both user-trigger analysis and session-trigger analysis, provided that the metric satisfies the additivity property.

3.3 Ratio Metrics

A Ratio metric has a form $X_i = Num_i/Denom_i$ where we use "num" and "denom" as shorthands for numerator and denominator. Both Num and Denom are additive, i.e., $Num = TrNum + UnTrNum$ and $Denom = TrDenom + UnTrDenom$. The complete potential outcome pair is

$$(Num_{iT}, Denom_{iT}, Num_{iC}, Denom_{iC}).$$

In the following discussion, we will assume $Denom_{iT} = Denom_{iC} = Denom_i$, i.e., there is no treatment effect on the denominator. This is an assumption that we need to provide a clean formula. We will discuss its rationale why it is a reasonable assumption. In fact in the next section, we will see in a unified framework we don't even need this assumption. Another key assumption is that there is no treatment effect on untriggered components, i.e. $UnTrX_{iT} - UnTrX_{iC} = 0$. Now we derive the formula for general ratio metrics. Readers are welcome to think about session success rate or click through rate. For session success rate, Num is successful session count and $Denom$ is session count. For click through rate, Num is click count and $Denom$ is page-view count. The following derivation also works for both user-trigger analysis and session-trigger analysis.

$$\begin{aligned}
&N \times \Delta_{overall}(X) = \sum(X_{iT} - X_{iC}) \\
&= \sum_{UTr}(X_{iT} - X_{iC}) + \sum_{Tr}(X_{iT} - X_{iC}) \\
&= \sum_{Tr}(X_{iT} - X_{iC}) \quad \text{no effect on untriggered users} \\
&= \sum_{Tr} \frac{Num_{iT} - Num_{iC}}{Denom_i} \quad \text{common denominator} \\
&= \sum_{Tr} \frac{TrDenom_i}{Denom_i} \times (TrX_{iT} - TrX_{iC}).
\end{aligned}$$

The last equation is from $Num_{iT} - Num_{iC} = TrNum_{iT} - TrNum_{iC} = TrDenom_i \times (TrX_{iT} - TrX_{iC})$.

$TrDenom_i/Denom_i$ is the trigger rate (TR) of the denominator. For session success rate, this is the session trigger rate for a triggered user. Divide both side by N we get the formula

$$\Delta_{Overall}(X) = \frac{1}{N} \sum_{Tr} TR_i \times (TrX_{iT} - TrX_{iC}) \quad (3)$$

If this trigger rate TR and the user level treatment effect $TrX_{iT} - TrX_{iC}$ are independent, the r.h.s. can be simplified

to[5]

$$1/N \times N_{Tr} \times \overline{TR} \times \overline{TrX_{iT} - TrX_{iC}}$$
$$= 1/N \times N_{Tr} \times \overline{TR} \times \Delta_{Tr}(X).$$

Therefore

$$\Delta_{Overall}(X) = \Delta_{Tr}(X) \times \frac{N_{Tr}}{N} \times \overline{TR}.$$

This is formula (2).

From this derivation, it is clear that formula (2) relies on the independence assumption which is normally not guaranteed. The counterexample we gave in the last section is a case where user 1 has both larger treatment effect and higher trigger rate, demonstrating a positive correlation between the two. In practice, there are many cases where a treatment will impact heavy user and light user differently and at the same time heavy user could also have higher or lower different trigger rate for a low coverage feature. The dependency between trigger rate and the treatment effect is very common. Therefore formula (2) is only an approximation that could be off when the aforementioned correlation is high.

3.4 Discussion: Implementation and Assumption Checking

Formula (3) cannot be simplified further into a direct translation from Δ_{Tr} to $\Delta_{Overall}$. For most cases, the trigger rate TR only has finite possibilities. If we first group all users by trigger rate,

$$\Delta_{Overall}(X) = \frac{1}{N}\sum_r r \times \sum_{TR=r}(TrX_{iT} - TrX_{iC})$$
$$= \frac{N_{Tr}}{N} \times \frac{1}{N_{Tr}} \times \sum_r r \times N_{TR=r} \times \Delta(TrX|TR=r) =$$
$$\frac{N_{Tr}}{N}\left(\frac{1}{N_{Tr}} \times \sum_r r \times N_{TR=r} \times \Delta(TrX|TR=r)\right).$$

The first term is still the user trigger rate. The rest in the parenthesis can be calculated by first grouping users by trigger rate TR, and then calculating the delta of the metric in each sub group, multiplied by trigger rate r. Those group level values are then averaged using the weight $N_{TR=r}/N_{Tr}$. Note that $N_{Tr} = \sum_r N_{TR=r}$ so these weights sum up to 1 and this last step is a weighted average.

In practice it is usually hard to match for all values of TR, we can resort to some discretization to reduce the number of sub groups. This also means we are sacrificing accuracy.

We do not recommend the approach above for all these practical concerns. The following is an alternative that is much easier to implement. Re-arrange r.h.s. of Formula (3) as

$$\frac{1}{N}\sum_{Tr}(TR_i \times TrX_{iT}) - \frac{1}{N}\sum_{Tr}(TR_i \times TrX_{iC}). \quad (4)$$

We see that to estimate the overall treatment effect of a metric X, instead of estimating its value in both treatment and control by taking sample average, we can conceive of a derived metric $Y_i = TR_i \times TrX_i$ and calculate sample average $\overline{Y}$ in treatment and control. Note that for untriggered users, $TR_i = 0$ and hence $Y_i = 0$.

$$\Delta_{Overall} = \overline{Y_{iT}} - \overline{Y_{iC}}.$$

This suggests that we only need to compute the Δ for this new dummy metrics to get an unbiased estimator for the overall treatment effect. This is considerably easier to implement than the trigger rate grouping method. In the next section we will further extend and improve this.

The assumption that there should be no difference for the trigger complement set is a more fundamental assumption that justifies using session-trigger analysis. If there is effect carried on from triggered session to untriggered session, then user-trigger analysis would be preferred. One way to check this assumption is to test this assumption by doing trigger-complement analysis, similar to how we test treatment effect in the trigger analysis. If there is no statistically significant movement, based on Occam's Razor principle, we'll happily accept this assumption.

Another assumption is that there is no effect on the denominator. This is also something we can test with statistical test. When this test fails, the idea of using this ratio metric is at stake. What does it mean if click through rate decreased but number of page-view increased? A feature can clearly make user happier to try our service more for harder tasks with lower success rate or conversion rate. For this reason, when we look at a ratio metric, it is fair to assume the denominator does not move. Otherwise we should avoid using this metric.

4. A UNIFIED FRAMEWORK: DILUTION AS VARIANCE REDUCTION

Formulas like (1) for additive metrics and (3) for ratio metrics leave a lot to be desired. First it feels redundant that we need different formulas for different type of metrics. What if we have another type of metrics that are not covered by either additive metrics or ratio metrics? The novel derivation method we displayed using potential outcome framework can be used on new types of metrics. But it won't guarantee the result can be simple enough to have an clean implementation. As we have already seen, implementating Formula (3) involves nontrivial investigation and we were lucky to have one.

Fortunately, there is a unified framework that solves this whole dilution problem in a more elegant way. The idea was laid out in *Improving the sensitivity of online controlled experiments by utilizing pre-experiment data* (1). The dilution problem has seemingly no relationship with pre-experiment data and yet the method is general enough to cast the dilution problem as a special case.

4.1 Dilution and Variance Reduction

The ultimate purpose of dilution is to more accurately estimate the overall treatment effect. If we also want to maintain unbiasedness of the estimator, by variance-bias tradeoff, reducing variance is the only way. The whole purpose of using trigger analysis and then applying correct dilution formula to translate the estimated effect back to overall population should be evaluated by two criteria. 1) Is the new estimator unbiased? 2) Is the variance of the estimator reduced?

[5]Rigorously, this is not exact equality. But they are both unbiased estimators for the same quantity under the independence assumption.

This line of thoughts lead us to a new approach. Why do we need to separate the job into two steps and try to derive the correct dilution formula? Can we tackle this problem as a whole and focus on reducing variance?

In Deng et al. (1), the authors provided a practical, almost assumption-less and intuitive approach to reduce variance of treatment effect estimator. The key step is to find other metrics that are not supposed to be affected by the treatment effect. These metrics are called covariates and are used to adjust the estimator similar to linear regression without all the burdens of linearity and normal assumptions of linear regression. Also see Yang and Tsiatis (7) and Tsiatis (6) for alternative approaches and general theory of semiparametric methods.

4.2 Variance Reduction Framework

Suppose X is the metric of interest and we wish to estimate the treatment effect

$$\mu = EX_T - EX_C.$$

The naive estimator is $\Delta(X) = \overline{X_T} - \overline{X_C}$. If we have another metric Y which is not affected by the treatment effect. Then

$$EY_T - EY_C = 0,$$

and therefore

$$\Delta^* = (\overline{X_T} - \overline{X_C}) - \theta \times (\overline{Y_T} - \overline{Y_C}) = \Delta(X) - \theta\Delta(Y) \quad (5)$$

remains to be an unbiased estimator for the same treatment effect μ for arbitrary choice of θ. Recall our goal is to reduce variance of the estimator. Minimizing the variance of Δ^* we get the optimal θ

$$\theta^* = \frac{\mathrm{Cov}(\Delta(X), \Delta(Y))}{\mathrm{Var}(\Delta(Y))} = \frac{\mathrm{Cov}(\overline{X_T}, \overline{Y_T}) + \mathrm{Cov}(\overline{X_C}, \overline{Y_C})}{\mathrm{Var}(\overline{Y_T}) + \mathrm{Var}(\overline{Y_C})}$$

where the second equality is due to the fact treatment and control are two independent groups of subjects. In practice since we don't need to use the exact optimal θ, we can simply use control group alone

$$\theta^* = \frac{\mathrm{Cov}(\overline{X_C}, \overline{Y_C})}{\mathrm{Var}(\overline{Y_C})} = \frac{\mathrm{Cov}(X_C, Y_C)}{\mathrm{Var}(Y_C)}$$

or similarly use treatment data, or form the average of both. When treatment effect is not huge, these difference choices do not make big difference. Under null hypothesis when there is no treatment effect, all those choices all converge to the same value. Also, as pointed out in Deng et al. (1), although it seems that we need to estimate θ^* from the data and hence a concern about accuracy of this estimation, it is really not an issue at all. This is because Δ^* is unbiased for any chosen θ and even if we don't accurately estimate the optimal θ^*, we still get most of the variance reduction from using a close enough estimate.

We can also see the optimal variance reduced from this approach.

$$\mathrm{Var}(\Delta^*) = (1 - \rho^2)\mathrm{Var}(\Delta),$$

where $\rho = Corr(X_C, Y_C) = Corr(X_T, Y_T)$. That is, the higher the correlation between X and Y the larger the variance reduced[6].

[6]In practice we don't use this formula to estimate the variance of Δ^*. Instead we directly work out the variance of (5) with estimated covariances and θ^* plugged in.

This method can apply to multiple covariates. When $\mathbf{Y}$ is multi-variate, θ^* is also a vector and the variance reduction is $1-R^2$ where R^2 is the variance explained by the covariates, similar to R^2 in multiple linear regression.

Deng et al. (1) showed that for online A/B testing, covariates Y are abundant. The authors recommended using the same metric of interest calculated using pre-experiment data. In their empirical study, these pre-experiment covariates can reduce variance for metrics like queries per user by 40% to 50%. The method was also called *CUPED* from initials in *Controlled experiments by Utilizing Pre-Experiment Data*.

4.3 Unified Dilution

In Section 3 we've derived different formulas for additive metrics and ratio metrics. From variance reduction perspective, we can unify all these into one step without differentiating additive and ratio metrics, and potentially support all different types of metrics. The implementation is also significantly cleaner and easier.

To apply variance reduction, we only need to propose covariates that are highly correlated to the metric of interest and not affected by treatment. The intuition behind the trigger analysis is that trigger complement data is not affected by the treatment effect and therefore only add noises. This makes it a perfect covariate! Note that we make no assumption on the form of metric X, be it additive or ratio or other functional forms.

Some other covariates available include trigger rate and other pre-experiment metric values. Pre-experiment metrics are the topics of Deng et al. (1) and we treat them as orthogonal to our focus in this paper. We propose the following variance reduction formula:

$$X \sim UnTrX + TR + (IsTR = 1), \quad (6)$$

where the r.h.s. lists covariates and the '+' means additional covariates. $UnTrX$ represents the same metric calculated from trigger complement data. For user with 100% trigger rate, there is no trigger-complement data so $UnTrX$ is not well defined. To deal with that, we add one additional binary covariate ($IsTR = 1$) that indicates whether there is no trigger complement data. When this is true, we define $UnTrX$ to be 0 or arbitrarily. See more discussion of this implementation detail in Deng et al. (1, Section 4.2).

In the last section we also showed for ratio metrics, sample average of a dummy metric $Y = TR \times TrX$ can be used to estimate overall treatment effect for a metric X. We can extend this by further applying variance reduction on top of it:

$$Y := (TR \times TrX) \sim UnTrX + TR + (IsTR = 1) \quad (7)$$

This is similar to Model (6) except the l.h.s. is replaced by dummy metric Y. The theory of variance reduction gurantees that this can produce another unbiased estimator for $EY_T - EY_C$, which is the same as $EX_T - EX_C$ based on Formula (3).

4.4 Connection to Dilution Formula

There is a deep connection between dilution with variance reduction framework and simple dilution formula. To see that, we use additive metrics as an example for simplicity and apply the VR formula $X \sim UnTrX$.

To apply variance reduction on additive metrics, since $X = TrX + UnTrX$, we get

$$\theta^* = \frac{\mathrm{Cov}(X, UnTrX)}{\mathrm{Var}(UnTrX)} = \frac{\mathrm{Cov}(TrX, UnTrX) + \mathrm{Cov}(UnTrX, UnTrX)}{\mathrm{Var}(UnTrX)} = 1 + \frac{\mathrm{Cov}(TrX, UnTrX)}{\mathrm{Var}(UnTrX)}$$

If we ignore the second term and $\theta^* = 1$, then the variance reduction formula (5) reduces to

$$\Delta^* = \Delta(X) - \Delta(UnTrX) = \frac{N_{Tr}}{N} \times \Delta_{Tr}(X)$$

since $X = TrX + UnTrX$ and $X = UnTrX$ for untriggered users and $\Delta_{Tr}(X) = \Delta_{Tr}(TrX)$.

The ignored second part of θ^* is trying to achieve additional variance reduction by exploiting the correlation between TrX and $UnTrX$. In other words, the dilution formula (1) is a special case with suboptimal choice of θ^* under the variance reduction framework, and can be superseded by variance reduction.

5. RESULTS AND DISCUSSION

We evaluate the performance of the unified dilution framework in two phases. High quality labeled A/B experiments were selected. Different models were applied to each metric and their variance reduction rates were compared. We further compared the two main trigger analysis methods: user-trigger analysis and session-trigger analysis.

5.1 Data and Metrics

Three A/B experiments that have different trigger rates are selected. The experiments were labeled positive or negative with high confidence. In each experiment, millions of users and queries are included with the rich click information. The scale of the data set is 1000 times more than what usually is affordable for human judgment data.

The evaluation is done for three methods: the formula approach that computes the delta for $TR \times TrSSR$ (implementation of Formula (3) in the form of (4) and we call it the Exact Formula), using $TR_i \times TrSSR_i$ as response (model (7)) and using SSR as response directly (model (6)) for the variance reduction framework. For all three methods, we can also compare user-trigger and session-trigger approaches. Among the three, only model (6) is generic and can be easily modified and applied to other types of metrics. Using $TR_i \times TrSSR_i$ either directly or as the response in model (7) are extensions of Formula (3) and can only be applied to ratio metrics.

For performance measures, we mainly focus on the variance reduction rate. The advantage of variance reduction rate is that it can be directly translated to the accuracy of the point estimation for the overall treatment effect, because higher variance reduction corresponds to better accuracy or narrower confidence interval of the estimate. Quite often, improvement of variance reduction also leads to a difference between whether the statistical test is significant or not. Variance reduction also translates to traffic or sample size saving. A x% variance reduction is equivalent to being able to reach to the same statistical power with x% less traffic/sample size.

Experiments	Trigger Rate	
	%Triggered Users	%Triggered Sessions
ExpA	5.26%	1.27%
ExpB	33.46%	20.83%
ExpC	65.17%	60.35%

Table 2: Trigger User and Trigger Session Ratios.

Experiments	User Trigger		
	Exact Formula	Mod. (7)	Mod. (6)
	VR rate	VR rate	VR rate
ExpA	88.60%	**98.42%**	95.60%
ExpB	-17.14%	**84.80%**	78.57%
ExpC	-49.47%	**61.45%**	36.03%

Table 3: User Trigger Three Models Variance Reduction comparison.

5.2 Variance Reduction Rate

We first characterize the trigger rate for 3 experiments **ExpA**, **ExpB**, **ExpC**. **ExpA** considers changing in positions in the Related Search blocks, **ExpB** focuses on video answer in task panes, while **ExpC** looks at creating dynamic sizing for results. Table 2 shows the percent of users and sessions funneled into user-trigger and session-trigger analysis for 3 example experiments. ExpA, for instance, have only 5.26% users exposed to the treatment and only 1.27% overall sessions have treatments in them.

We have carried out another set of 3 experiments with triggering rates close to **ExpA**, **ExpB**, **ExpC** respectively. The results for both sets are similar. For pair of flights with very different treatments, when trigger rate is similar, the resulting variance reduction rate is similar. we have done analysis for dozens of experiments with other trigger rate values. As expected, trigger rate is the key factor that affects the variance reduction rate while the particular feature changes do not matter much.

ExpA, **ExpB**, **ExpC** are also different types of flights in themselves. **ExpA** considers changing in positions in the Related Search blocks, **ExpB** focuses on video answer in task panes, while **ExpC** looks at creating dynamic sizing for results. The trigger rates for the 3 experiments are illustrated in the Trigger Rate Table. The number of total users ranges from 28 million to 40 million for the 3 experiments.

Table 3 shows the results on the 3 experiments for user-trigger models using Exact Formula (4), variance reduction method using model (7) and model (6). VR method using model (7) yields the largest variance reduction rate in all three experiments. It is surprising that using the exact formula directly without VR can lead to increased variance for high coverage features as shown in ExpB and ExpC. Turned out this is due to additional variance contributed by the multiplication factor TR. For high coverage feature where TR can be large for some user but low for others, the variance in TR can generate extra variance. To get most out of the exact formula, we need to group users by their trigger rate TR to remove variance contributed by TR. This is done implicitly by adding TR as one of the covariates in model (7), which eventually leads to the largest variance reduction. However, even without any cue from the exact formula (4), using the generic variance reduction model (6) also performs quite well.

Experiments	Session Trigger		
	Exact Formula	Mod. (7)	Mod. (6)
	VR rate	VR rate	VR rate
ExpA	97.12%	**99.44%**	98.25%
ExpB	28.12%	**89.85%**	85.99%
ExpC	-31.32%	**69.10%**	53.97%

Table 4: Session Trigger Three Models Variance Reduction Comparison.

Experiments	User Trigger	Session Trigger
	Variance	Variance
ExpA	4.40%	1.75%
ExpB	21.43%	14.01%
ExpC	63.97%	46.03%

Table 5: Variance Using Mod. (6): User Trigger, Session Trigger. Variance is measured as percentage of the variance from all up analysis.

Table 4 compares the three models using session-trigger covariates for the two VR methods. Similar to user-trigger scenario, VR method using $TR \times TrSSR$ (model (7)) has the highest variance reduction rate in all 3 experiments even if directly using $TR \times TrSSR$ can increase variance as in ExpC.

We now compare the variance reduction rate for user-trigger and session-trigger methods. We use VR method with SSR (model (6)) in the example because this is a generic method not depending on any theoretically derived formula. Table 5 shows the variance using Model (6) measured as percentage of corresponding original all up analyses' variances. Session-trigger method consistently leads to smaller varaince. Note that although the varaince reduction rate in previous tables might not be quite different for low coverage experiment as in ExpA, the true variance difference is huge. To see that, for ExpA, the variance after reduction is 4.4% (user-trigger) compared to 1.75% (session-trigger). Session-trigger with VR model (6) produced a variance that is only 40% of the variance from user-trigger with the same VR model. This also clearly translates to the sensitivity improvement that reflects in the p value.

For all 3 experiments, the change from using user-trigger to session-trigger leads to the improvement for making correct data-driven decisions. In both ExpB and ExpC, user-trigger analysis are boarder line significant, which makes it hard to draw a clear decision by itself. Session-trigger analysis, however, confirms the movement and lifts the confidence of the decision, a clean data-driven decision can be drawn directly.

5.3 Impact on Hypothesis Testing

If we are not concerned about the "value centric" estimation for relative delta, but rather only focus on the "action centric" hypothesis testing, then it is possible to do A/B experiments only based on user-triggered sessions or session-triggered sessions, i.e., the analysis methods shown in Figure 2, also known as trigger analysis. Table 6 compares the t-statistics for using generic VR method model (6) to using trigger analysis. They both gave statistically significant results in the 3 experiments. VR methods won on ExpA and the trigger analysis won for ExpB. Note that t-statistics

Experiments	VR Using SSR	Trigger Analysis
	Session Triggered	Session Triggered
	t stats	t stats
ExpA	6.72	3.32
ExpB	3.97	5.26
ExpC	3.50	3.39

Table 6: Sensitivity comparison for VR model using SSR and trigger analysis.

Session Trigger Complement			
Experiments	ExpA	ExpB	ExpC
P value	0.23	0.38	0.55

Table 7: P value for A/B testing on the session trigger complement.

are random numbers influenced by the estimated delta. In connection to the discussion in Section 4.4, we believe in general VR metric would be slightly more powerful than trigger analysis due to its extra variance reduction. But the main advantage of using VR method is the direct estimation of overall treatment effect, not just the treatment effect on triggered population.

5.4 Assumption Checking and Final Remarks

Results in this section suggest that we can use the variance reduction model directly for estimating overall treatment effect without the two steps approach involving a trigger analysis and a dilution step. For a general metric X, we recommend model (6) which can be easily implemented. For ratio metrics, model (7) is a strong alternative. However it does rely on extra assumption such as no treatment effect on the denominator of the ratio metrics which is used to prove formula (3). This makes it less reliable and applicable.

Another assumption for doing session-trigger analysis is that the treatment effects are contained completely in triggered sessions. One way to verify this assumption is to do hypothesis testing for the metric of interest on session-trigger complement sets. Table 7 shows that all the resulting p values are not statistically significant for the three experiments. Therefore, we don't see evidence there are any treatment effects in session-trigger complement sets in these three examples[7].

6. CONCLUSIONS

In this work we looked at the dilution problem for estimating the overall treatment effect in online controlled experiments. We showed the right way for doing this is not obvious for a large set of widely used metrics, the ratio metrics. We first theoretically derived the dilution formula, based on Rubin's potential outcome framework. Furthermore, we discovered the connection between the variance reduction framework and the dilution problem and successfully reformulated the problem as a special case of variance reduction. We presented empirical results for the unified dilution appro-

[7]Note that we can never prove there is no effect in session-trigger complement, unlike user-trigger compliment. And any statistical test have false negatives. We call out Occam's Razor in this case that we choose to believe a simpler model(no effect in session-trigger compliment) given no evidence against it.

ach and demonstrated that it yields significant improvement in the accuracy for the overall treatment effect estimation. We compared two different trigger analyses, session-trigger analysis and user-trigger analysis in detail. We are not aware of other literature on the same topic. We believe the work here is a novel application of the variance reduction framework in "value centric" online controlled experiments and greatly improves the accuracy of treatment effect estimation for low coverage features.

7. ACKNOWLEDGMENTS

We wish to thank Ron Kohavi, Aron Inger, Toby Walker, Brian Frasca and others from Microsoft Analysis and Experimentation Platform, Georg Buscher, Aidan Crook, Siamak Faridani, Widad Machmochi, Anand Oka from Bing Measurement group and Jan Pedersen from Bing Search Relevance for providing the problem, motivation, suggestions, comments and support on this work. We also wish to thank Yu Guo, Xiaolin Shi, Roger Longbotham for reading early draft and giving thoughtful feedbacks. Last but not the least, we thank all 4 reviewers who gave valuable suggestions to help us improved this paper.

APPENDIX

Illustrative Toy Example

In this section we give detailed steps of 2 methods we compared in Section 5: (1) the formula approach that computes the delta for $TR \times TrX$ (Formula (4)), (2) using X as response directly (model (6)). Model 7 is very similar to Model 6 so we leave it as an exercise to readers. As in Section 5 we chose the metric X to be SSR. Note that the framework in this paper assumes enough sample size for central limit theorem to be applicable. Small sample toy example like this is only for illustration.

Group	User	S1	S2	S3	S4	S5	X	TR	TrX	UnTrX	TR=1
T	A	1	***0***	0	0	1	2/5	1/5	0	1/2	0
T	B	***1***	***1***	***0***	***1***		3/4	1	3/4	0	1
T	C	***1***	0	0			1/3	1/3	1	0	0
T	D	0	0	0			0	0	0	0	0
C	E	0	1	***0***	1	1	3/5	1/5	0	3/4	0
C	F	***1***	***1***	***1***			1	1	1	0	1
C	G	0	0	1			1/3	0	0	1/3	0
C	H	0	***1***	0	0		1/4	1/4	1	0	0

We use the toy example listed in the table above throughout this section. 8 users in this example are split into treatment and control with up to 5 sessions observed. Each entry in the table indicates success or not, and triggered sessions are emphasized as bold and italic.

The first step is to calculate the metric value X, the trigger rate TR and the trigger-complement metric value $UnTrX$. In our toy example of SSR, X is simply the average of each data row, and trigger rate TR is the number of bold and italic entry divided by the number of non-empty entries. $UnTrX$ is the SSR calculated using plain entries(not bold and italic). Two special cases are user B and F, who have all sessions triggered. Hence there is no trigger-complement sessions and we define $UnTrX$ as 0 and also mark $TR = 1$ as 1.

Exact Formula 4. Formula 4 is straightforward to apply. We just take delta of $TR \times TrX$ calculated in treatment group and control group.

$$\text{T} : (1/5 \times 0 + 1 \times 3/4 + 1/3 \times 1 + 0)/4 = 0.271$$
$$\text{C} : (1/5 \times 0 + 1 \times 1 + 0 \times 1/3 + 1/4 \times 1)/4 = 0.313$$
$$\text{Overall Effect} : 0.271 - 0.313 = -0.042$$

For test, we calculated variance of $TR \times TrX$ for the two groups to be 0.127 and 0.224. And the variance of the delta is estimated to be $0.127/4 + 0.224/4 = 0.088$. Finally we get our z-score of $-0.042/\sqrt{0.088} = -0.142$.

VR: Formula 6. To apply Formula 6 the gist is to estimate θ^*. Note that $\theta^* = \text{Var}(Y)^{-1} \times \text{Cov}(X, Y)$ where Y is the vector $(UnTrX, TR, IsTR = 1)$ and Var and Cov here are matrices. For the control group, θ^* is

$$\begin{pmatrix} 0.127 & -0.081 & -0.090 \\ -0.081 & 0.192 & 0.213 \\ -0.090 & .213 & 0.250 \end{pmatrix}^{-1} \times \begin{pmatrix} -0.010 \\ 0.130 \\ 0.151 \end{pmatrix} = \begin{pmatrix} 0.488 \\ 0.317 \\ 0.512 \end{pmatrix}.$$

For illustration we'll just take this $\theta*$. In practice, one can use a weighted average of θ^* estimated from control and treatment. Using this θ^*,

$$\begin{aligned} &\Delta_{VR} = \Delta(X) - 0.488 \times \Delta(UnTrX) \\ &- 0.317 \times \Delta(TR) - 0.512 \times \Delta(IsTR = 1) \\ &= -0.175 - 0.488 \times (-0.145) - 0.317 \times 0.021 - 0.512 \times 0 \\ &= -0.111 \end{aligned}$$

To get z-score, we also need to calculate the variance. which is

$$\begin{aligned} &\text{Var}(\overline{X}_T) + \text{Var}(\overline{X}_C) + (\theta^*)^T \left(\text{Cov}(\overline{Y}_T) + \text{Cov}(\overline{Y}_C)\right) \theta^* \\ &- 2 \times (\theta^*)^T \left(\text{Cov}(\overline{X}, \overline{Y}_T) + \text{Cov}(\overline{X}, \overline{Y}_C)\right) \\ &= 0.00435 \end{aligned}$$

The variance reduction rate hence is $1 - 0.00435/0.086 = 95.0\%$. Z-score is then $-0.111/\sqrt{0.00434} = -1.685$.

References

[1] Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 123–132. ACM, 2013.

[2] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining Knowledge Discovery*, 18:140–181, 2009.

[3] Ron Kohavi, Alex Deng, Roger Longbotham, and Ya Xu. Seven rules of thumb for web site experimenters. 2014.

[4] Jim Manzi. *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society*. Basic Books, 2012.

[5] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[6] Anastasios Tsiatis. *Semiparametric Theory and Missing Data*. Springer-Verlag, 2006.

[7] Li Yang and Anastasios A. Tsiatis. Efficiency study of estimators for a treatment effect in a pretest-posttest trial. *The American Statistician*, 55, 2001.