Estimating True Ad Lift in Randomized Tests with Auction-Winning Bias

March 2, 2016

1 Background

This note describes a methodology for estimating true lift of a digital advertising campaign from observed response rates in a randomized experiment. The scenario we consider is as follows. When bid-requests arrive at the DSP, the bid-opportunity is matched against eligible strategies, and various targeting and other filtering actions are performed, to identify a campaign that will bid on this opportunity. But *before* the DSP bids on the opportunity, the cookie in the bid-opportunity is randomly assigned to either *control* (with 10% probability) or *test*. Note that test/control assignment depends only on the cookie. If the cookie is assigned to control, the corresponding bid-opportunity receives no bid from the DSP for this campaign, hence the user will not be exposed to this campaign's ad via this DSP (however the user may of course see ads from the same or other advertiser via other platforms, or even from other advertisers from this DSP).

If every test cookie gets to see an ad from the campaign, then we would have a reasonably close approximation to a clean randomized experiment to estimate true lift: simply observe response rates of the test and control populations, and compute lift. However, the DSP is only guaranteed to *submit* a bid for the opportunity assigned to test, and *there is no guarantee it will win the auction*. This is the key complication we address in this note:

Before submitting the bid, the test and control populations are nominally equivalent, since the test/control assignment is done after all targeting and matching; however the test *winner* population is most likely *not* equivalent to the control population. For example, the DSP may be more likely to win impressions for which there is less competition in the auction, and hence could be of lower quality. Thus if we imagine there are high-responders ("gold") and low-responders ("rocks") in the control population, the test-winner population could be skewed toward rocks. In other words there is a potentially large *win-bias* in the test population that is exposed to ads.

2 Notation

We use R to denote the response-rate, combined with superscripts and subscripts. Superscripts indicate ad-exposure: 1 means exposed, 0 means not exposed, Subscripts denote the population. Certain combinations of population and ad-exposure are "impossible"; in these cases we are referring to the response-rate in a *hypothetical* (or *counterfactual*) scenario where we imagine that combination to occur.

Specifically, R^1_{tw} denotes response rate of test-group winners when exposed to an ad (we know all test-group winners are exposed to ads, but the 1 serves to make this explicit). R^0_{tw} is the response rate of the test-group winner, *if* they were *not* exposed to an ad. This is clearly a counterfactual scenario. Similarly define R^1_{tl} and R^0_{tl} for the test-group losers.

Also let R^0_t, R^1_t denote the overall (counterfactual) response rates for the test group in the absence of an ad-auction, if none and all were exposed to ads, respectively.

For the control group there is no ad-exposure, so we just use R^0_c to denote the response-rate of a control user.

3 Formulation

Since test and control group assignments are random, and occur *after* targeting and matching, we can assert that:

$$R_t^0 = R_c^0, \tag{1}$$

where we note that R_c^0 is directly observable, but R_t^0 is a counterfactual: it is the (hypothetical) response rate of a test user *not* exposed to an ad; because of the auction dynamics there *will* be test users not exposed to ads, however R_t^0 is *not* simply the average response rate of the un-exposed test users, because of selection bias introduced by the auction process. In this sense R_t^0 is not *directly* observable.

Our goal is to measure the impact of an ad on a user. An initial naive approach to do this might be to compare these two *observable* response rates:

- R_{tw}^1, the response-rate of test-group winners (all of whom would see an ad), and
- R_c^0, the response-rate of the entire control group (none of whom see an ad).

This would in general not yield a true measure of lift, due to the win-bias issue mentioned earlier: even though the test and control populations are equivalent (from the point of view of conversion-rates) pre-auction, the test *winners* group is likely to have significant selection-bias.

The *true lift* we really want to measure is:

$$L_{true} = (R_{tw}^1 \ / \ R_{tw}^0) - 1, \tag{2}$$

i.e., how much higher is the test group winner's conversion rate when *exposed* to an ad, compared to their response rate *if they had not been exposed* to

an ad. This is referred to as the *Average Treatment Effected on the Treated* (ATET), in the econometrics literature. In the above naive aproach we have been implicitly using fact (1) combined with a tacit assumption that test winners and losers are equivalent in their (un-exposed) response rates, i.e., $R^0_{tw} = R^0_{tl} = R^0_t$, and using $R^0_c = R^0_t$ as a proxy for R^0_{tw} in eq (2). In general however R^0_c may not be a good proxy for R^0_{tw}. Indeed in several experiments we have seen that the test-winner conversion-rate R^1_{tw} is smaller than the control conversion-rate R^0_c.

The difficulty with using the true-lift formula (2) though is that the counter-factual R^0_{tw} is not directly observable. Below we show how this hypothetical response-rate of *unexposed test-group winners* can be derived from the observed R^0_{tl}, which allows us to compute true lift. The aim of this note is to derive a formula for true lift based on the conversion-rate of unexposed test-losers, and also to quantify how large of an effect to expect.

4 Derivation

Note that if w is the win-rate in the test-group (i.e. a fraction w of the test-group actually see an ad), then we can say:

$$R^0_t = wR^0_{tw} + (1-w)R^0_{tl}, \tag{3}$$

where R^0_{tl}, the response rate of test-group losers (under no-ad exposure), and is directly observable **if we log losses**. Also we noted in (1) that $R^0_t = R^0_c$ so we solve (3) for the hypothetical R^0_{tw} to get:

$$\begin{aligned} R^0_{tw} &= [R^0_t - (1-w)R^0_{tl}]/w &(4)\\ &= [R^0_c - (1-w)R^0_{tl}]/w. &(5)\end{aligned}$$

Now we can use (2) to compute true lift in terms of observable quantities:

$$L_{true} = \frac{wR^1_{tw}}{R^0_c - (1-w)R^0_{tl}} - 1. \tag{6}$$

It will be useful to consider the *ratios* $F_w = R^0_{tw}/R^0_c$ and $F_l = R^0_{tl}/R^0_c$. These represent the unexposed test-winner and test-loser response rates in units of the control response-rates. In terms of these ratios we re-write (5) as:

$$F_w = [1 - (1 - w)F_l]/w, \tag{7}$$

so in terms of the naive lift $L_n = R^1_{tw}/R^0_c - 1$ and the ratio F_l we can write the true lift formula as:

$$\begin{aligned} L_{true} &= (R^1_{tw} \ / \ R^0_{tw}) - 1 && (8)\\ &= (R^1_{tw}/R^0_c) \ / \ (R^0_{tw}/R^0_c) - 1 && (9)\\ &= (1 + L_n) \ / \ F_w - 1 && (10)\\ &= \frac{w(1 + L_n)}{1 - (1 - w)F_l} && (11) \end{aligned}$$

If we ignore win-bias, i.e. assume that $R^0_c = R^0_t = R^0_{tl}$, then we have $F_l = 1$ and this formula reduces to our original naive lift estimate (**??**).

To get a better sense of the factor F_l, solve above to get

$$F_l = \left(1 - \frac{w(1 + L_n)}{(1 + L_{true})}\right) \frac{1}{1 - w}, \tag{12}$$

This form allows us to answer how much larger must the unexposed test-loser response-rate be than the control conversion rate, if we want to turn a naive negative lift into a true positive lift? For example for $L_n = -20\%$ and $w = 5\%$, and $L_{true} = 5\%$, the factor F_l is

$$\begin{aligned} F_l &= (1/0.95) * (1 - 0.05 * (1 - 0.20)/(1 + 0.05)) && (13)\\ &= 1.0125, && (14) \end{aligned}$$

i.e. the test-loser (unexposed) response rate is around 1.25% higher than the control conversion rate. If we repeat the above calculation with $L_{true} = 20\%$, we get $F_l = 1.0175$, i.e. to turn a naive negative lift of -20% into a true positive 20% lift, the test-loser conversion rate need only be 1.75% larger than the control conversion rate. This gives a rough idea of how much effect we should be looking for, when trying to compute true lift by logging losing bids.