ANALYSIS OF BINARY OUTCOME A/B TESTS - NEW METRICS

Ramesh Subramonian, Ranjeet Tate, Michael Shire, Abhi Singh

As a user (Product Manager, Analyst, HiPPO) of AB tests or a designer/builder of AB tests and systems, do you have a nagging feeling that though A was statistically better than B, the difference wasn't large enough to be *important*? Do you have doubts about the *business impact* of the test outcome? Do you want to set a single-valued business or revenue goal (e.g. 10% better) and have a *definite recommendation* made to you instead of dealing with "we are 90% confident that A is 11.1% better than B"? Do you feel overwhelmed by the *"test everything"* approach becoming increasingly prevalent in Tech? If you answered yes to any of these questions, then read this blog.

First, we need to go beyond asking simply whether "A is better than B" to asking whether "the difference between A and B is important enough to be actionable". Consider an example: The Optimizely book on A/B testing describes a test comparing a page with a static ad to one with a video[1]. The variant with the video was better in terms of both click-through and conversion rates, but the costs of producing the video and displaying it were deemed "too high" to launch the video version at scale. So in addition to being statistically significant, the difference has to be *important* enough to recommend proceeding[2].

Second, this difference has to be evaluated not in terms of the usual probability but in terms of a revenue or business metric, which may not be simply related to the measured probability. E.g., in order to trigger a recommendation, do we want the *difference* between success probabilities to exceed 0.1 —i.e. $p_A - p_B > 0.1$— or do we want the success probability to *lift* by 20% —i.e. $p_A > 1.2 * p_B$? Different metrics will lead to different outcomes. If p_B happens to be 0.5, even though $p_A = 0.6$ has the same value for both a difference of 0.1 and a lift of 20%, the confidence levels associated with the two propositions can be quite different.[3]

Third, decision makers do not want to deal with statements about "confidence levels in the difference". The *goal of our analysis is to provide the test owner with a definite recommendation about choosing A or B*[4] based on a single "acceptance value" for a pre-selected business metric.

As a collateral benefit, you may find that the requirement of calculating a minimum business gain *before* starting a test will magically reduce the number of tests, culling out the more frivolous ones.

[1] "Fail Fast and Learn", pg. 79, *A/B Testing*, Dan Siroker and Pete Koomen, Wiley (2013)

[2] See `http://shopperscientist.com/archive/views/28may95.html` and `http://www.med.uottawa.ca/sim/data/Statistical_significance_importance_e.htm` for lengthier discussions.

[3] Note that the choice of comparison metric is an issue that arises *only* when we want to quantify the comparison. If we were only interested in *whether* A is better than B, then any metric (as long as it is monotonic in p_A and p_B) would do. The choice of metric becomes important when we are interested not just in whether A is better than B, but in addition, *by how much*.

[4] Paraphrasing from Klugman *et al* "Loss Models", 2nd Ed. Wiley (2004), pg. 419: "...the process must end with a winner. While qualifications, caveats etc. are often necessary, a commitment is required." on the part of the analyst.

We will illustrate our approach to the first three issues by analysing the data for binary outcome A/B tests, which we briefly describe. In a *binary* experiment, each trial has only two possible outcomes, a *success* (1) or a *failure* (0). A population being experimented on has a "true" or intrinsic probability of success p which is to be inferred from the results of the test. In an A/B test there are two populations or branches, each of which is exposed to a different treatment. The populations have intrinsic success probabilities p_A and p_B. After a period of time (pre-determined by "power" calculations or otherwise), the experiment will yield a count of the trials and successes in each branch, (n_A, m_A) and (n_B, m_B). From this data we infer a function that represents the likelihood of (p_A, p_B).

We've taken a Bayesian approach since it turns out to be much better suited for general metrics. We use the data from the experiment to construct the posterior probability distribution (or *likelihood*) of (p_A, p_B). Fig. 1 is the two dimensional space of intrinsic probabilities $\{(p_A, p_B)\}$, on which

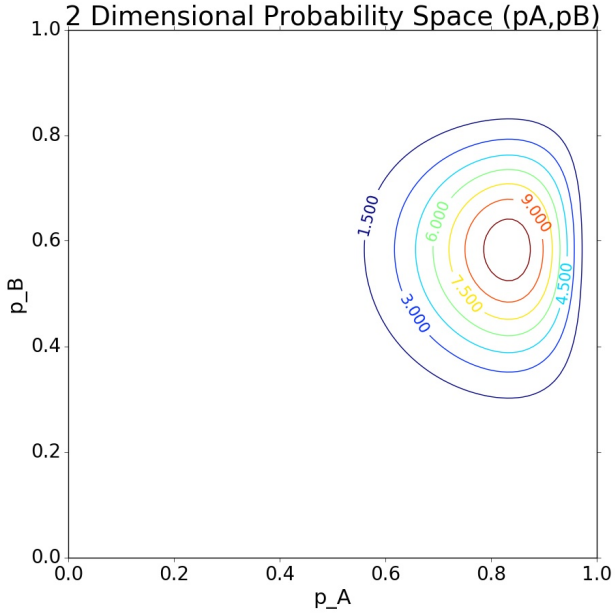

Figure 1: Contour Plot of the Likelihood of (p_A, p_B) for $(n_A, m_A, n_B, m_B) = (12, 10, 12, 7)$. The contours correspond to points of equal likelihood. The insides of high-value contour lines represent regions of high likelihood, with the peak at $(10/12, 7/12)$.

we've shown a contour plot of the likelihood function, see the caption for additional explanation.

The next issue is that of finding a metric which reflects a business interest. Consider a Binary A/B test in which the probability being measured is the *page transition probability* p: that of moving to any other page on the website as opposed to leaving the website or otherwise ending the session. The business impact is in fact *not* proportional to an increase in the page transition probability. We

know that on average revenue is proportional to clicks, conversions or other monetizing actions, and that the number of such actions is proportional to the number of opportunties to act, which in turn is proportional to the pages visited on the website. How is the number of page visits PV related to the page transition probability p? Some thought shows that the average number of pages visited per user-session is

$$PV = p + p^2 + p^3 + ... = \frac{p}{1-p} \tag{1}$$

which is simply the *odds ratio* corresponding to the probability[5]!

The same metric also occurs (surprisingly) in the context of loans: For a lender, the expected return on a loan is proportional to the number of loan payments made before a default. If p (related to the FICO score) is the probability of making any single loan payment, then Eq. 1 represents the average number of payments before default. Since p is close to 1, even a small increase in p leads to a large increase in expected return on the loan, and a corresponding drop in the APR the lender can afford to offer.

Different metrics (and their values) define different lines in two-dimensional probability space $\{(p_A, p_B)\}$. Fig. 2 is 2D probability space as in Fig. 1 where in addition to the contours of the likelihood we have plotted the lines defined by a metric value for each of three metrics: Probability Difference, Probability Lift and Page Views Lift.

If we were to know p_A, p_B with certainty, it would be a point in the above probability space. We would then recommend A over B if the observed value for the metric exceeded the chosen minimum value —equivalently, if the observed point lies below and to the right of the chosen metric line.

Note however that we do *not* have a point (p_A, p_B) that corresponds to our knowledge of the success probabilities, instead we have the likelihood function, which we have superposed on the metric lines in the figure above. Thus, we can ask for the (total) likelihood that the metric exceeds the value M, equivalently, that (p_A, p_B) lies below (and to the right of) the metric line for M. This is simply the volume of the likelihood function below the metric line for M, and is the *credibility* that the metric exceeds M.

(As we can see from the figure, the different metric lines cut the likelihood function on different sides of the peak, and so the calculated credibilities will be different and the resulting recommendations can be contradictory.)

So for any value of the metric M we can do a numerical integration to obtain its credibility $Cred(M)$[6]. For concreteness, consider experimental results $(n_A, m_A) = (200, 40)$ and $(n_B, m_B) =$

[5] More on this in a separate blog.

[6] We analytically reduce the double integral involved to 1 dimension, and then do a cute trick that allows for an easy numerical integration.

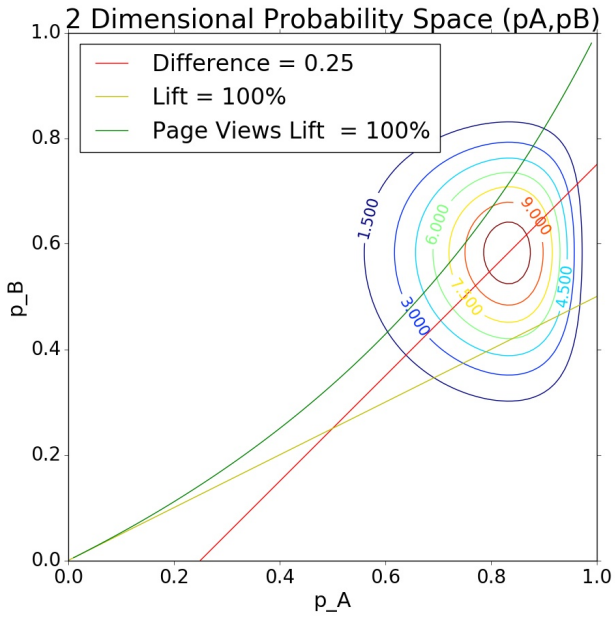

Figure 2: Metric Lines and Likelihood Function on 2D probability Space (p_A, p_B). The credibility of a metric is the volume of the likelihood function *below* the metric line.

$(100, 15)$. The analysis we've described so far provides a *Page Views Lift vs. Credibility* curve based on the experimental data, of the form in Figure 3. which as expected is sigmoidal. Thus, we

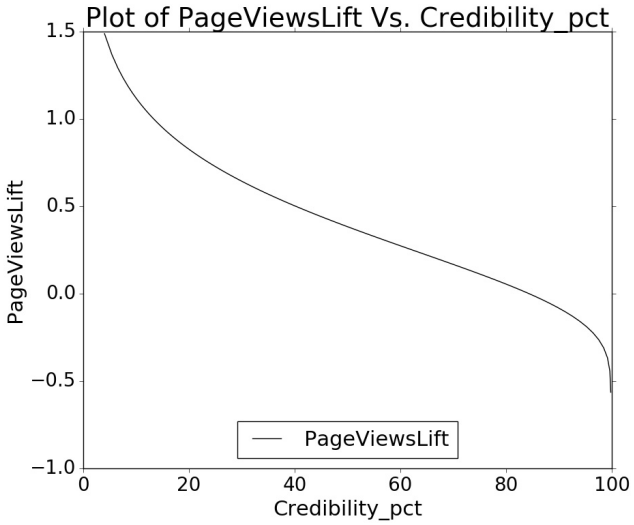

Figure 3: Page Views Lift vs. credibility

use the data to infer not just *which treatment is better*, but in addition *by how much* and *how certain*

we are of this. In the conventional approach, we would base our recommendation on a decision criterion like "Page Views Lift of 8% at 95% credibility", and if the point $(95, 0.08)$ is below and to the left of the above credibility curve in Figure 3 we would recommend A over B.

However, from the perspective of an *expected* business return, this process is ambiguous. Specifically, the above decision point lies above the curve and so "A is not better than B". But the decsiion point also corresponds to an expected return of $0.08 * 0.95 = 0.075$, which is equivalent to a "Page Views Lift of 10% at 75% credibility", which *does* lie below the credibility curve and thus implies a recommendation of "A is better than B". This still does not help either us or the HiPPO, who furthermore wants to express her or his decision criterion as a single value of expected return.

Our approach to resolving this conundrum is to interpret $M * Cred(M)$ as the *expected minimum value* of M. We plot this as a function of the credibility in Figure 4. This expected minimum value

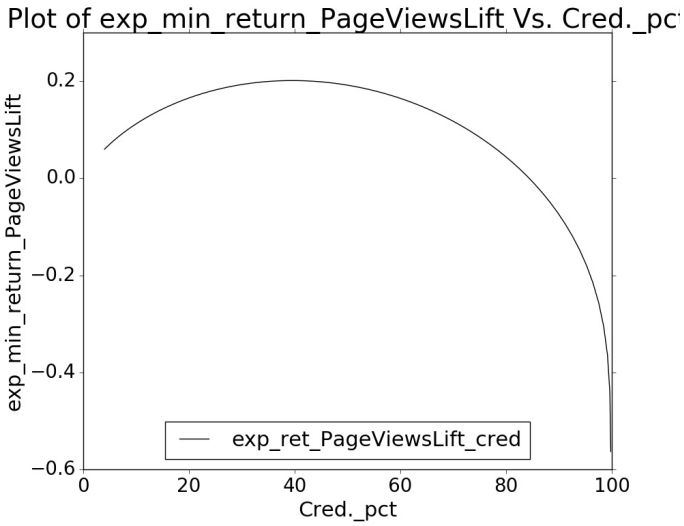

Figure 4: Expected Minimum Page Views Lift vs. Credibility

has a maximum. The question of what recommendation to make is then reduced to comparing the experimentally determined *maximum expected minimum value* to the test owner's single value M_{Min} for an expected return: if the *max-min* is lower than the M_{Min} then the variant is *not* better than the control. Conversely,

$$\text{If Max(Expected Minimum}M) > M_{Min}, \text{ then we recommend A over B.}$$

Let's recapitulate this last part. The test owner has selected a metric (let's say Page Views Lift) and a threshold value M_{Min} by which A has to exceed B in order for the results to be called in favor of A. From the data, for every value of the metric M we can calculate the credibility (Fig. 3). Note that $Cred(M)$ is the credibility that the *minimum* true value of the metric is M. We multiplied M

by its credibility, this is the *expected minimum* value of the metric. As a function of the credibility (Fig. 4), this expected minimum itself has a *maximum*. Clearly, if the *maximum expected minimum value* is less than the minimum acceptable value M_{Min}, we *cannot* call the results in favor of A. If the opposite holds, we *choose* to call the results in favor of A.

To summarize, we've shown two reasons to go beyond the "Is A better than B?" approach to AB testing: One, we have to take the importance of the difference between A and B into account, and two, we have to quantify the difference in terms of a business goal. Further, our approach allows the test owner to establish a single metric and value as a decision criterion, then our analysis provides a simple "A/B" recommendation, without any "confidence levels" clouding the issue.