White paper

Quantify: Decision making, multivariate testing and data exploration via robust statistical modelling

www.admetrics.io

admetrics

1. What is Quantify?

Quantify takes a new approach to making core marketing performance metrics credible and actionable. It automatically adds robust statistical models which

- **qualify** performance metrics for each **variation** by adding **credible ranges** based on the amount of available data, and
- **directly model performance lifts** among variations, reporting likely lifts and credible ranges; this allows easy filtering of practically significant effects

Without the need for manual interventions by dedicated experts, decision makers can now easily

- track and drill into ongoing campaigns and tests, and
- retrospectively explore past traffic for insights.

Our approach sidesteps the limitations of classical ("A/B") significance testing, delivering actionable insights to the business faster.

2. Quantify by practical example

Let's take a look at data of an advertiser in November 2018[2]. The two key performance metrics for a click-based attribution model are

- vCTR%[3] - viewable click-through rate, as percentage of viewable impressions
- CTC‰ - click-through conversion rate, as per-thousand of clicks

Conventional reporting tools usually show simple averages by variation for these metrics, while Quantify enables credible ranges based on the amount of available data, such that with 95% probability, the "true" value of the metric is contained in the credible range (see Figure 1)

The width of a credible range depends on the amount of data available. For instance, vCTR% will typically have tighter credible range bounds than CTC‰, since the data set of viewable impressions is far larger than the data set of conversions.

Since we are interested in comparisons between variations, we also directly model the relative lift in performance[4].

---

[1] – splits of the traffic, see below for more

[2] – Data has been anonymized

[3] – One might argue that vVTC‰ (view-through conversion rate as per-ten-thousand of viewable impressions) is a better metric, but these campaigns reported on vCTR%.

[4] – Underlying math explained later in this document

Deriving lift requires picking one variation as baseline. We refer to all other variations as challengers. In an A/B test, A = baseline, and B = challenger.

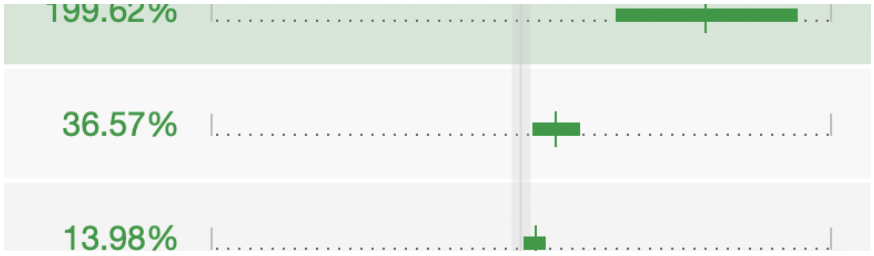


Figure 1: A lift visualized as credible range through a bar. The most probable lift is marked vertically in the bar at 22.66%.

In our tour, our experiment looks at vCTR for 300x250 creatives. The variations are:

- two larger publishers
  - MultiMeter
  - Sloth
- two smaller publishers with lower traffic numbers
  - Rightbutton
  - Flower

First, for a single day[5] of data , we look at our four publishers. The explorer view outputs counts and averages as shown in Figure 2. But how certain can we be about the relative performances of the remaining trio?

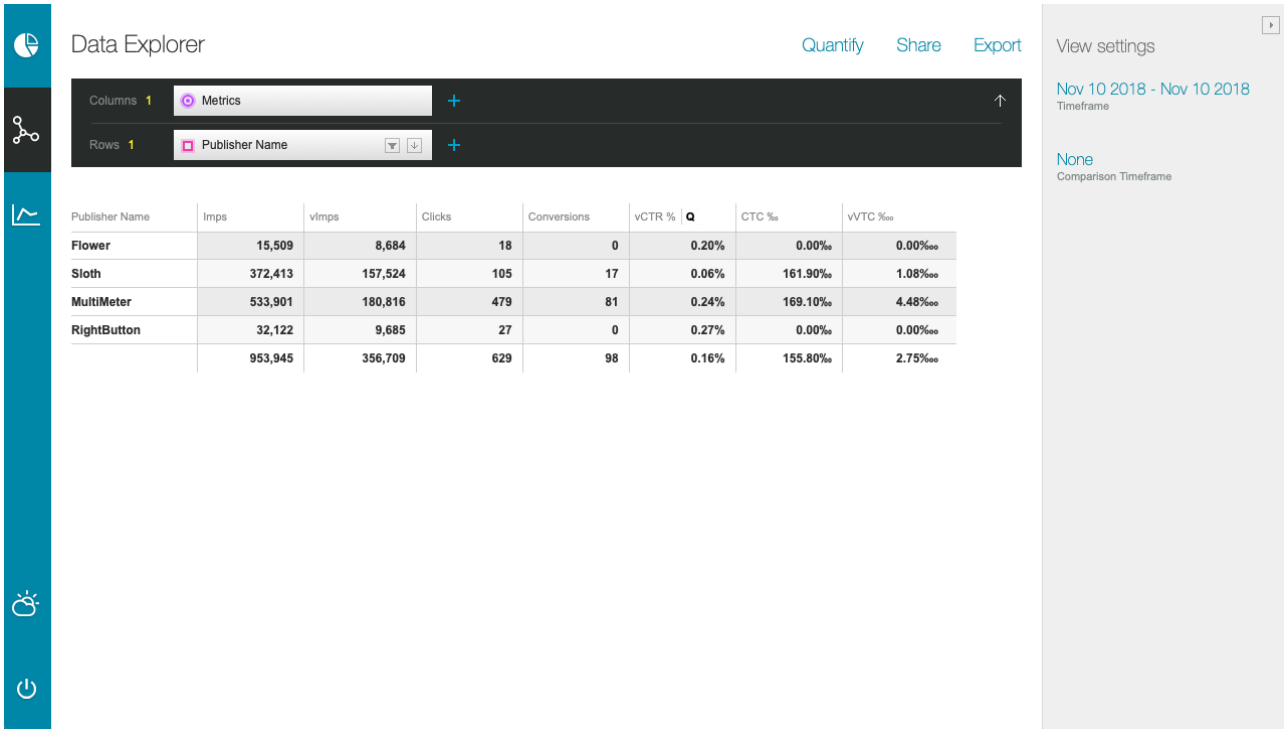
Engaging Quantify

We compare all challengers against our adaptive debiased average (see section 4 for more details on how this is calculated) (Figure 3)

We see that ***RightButton***, one of the lower traffic challengers, may be better, while ***Sloth***, the other, is worse. ***Flower*** completely overlaps the effect band, which is set at its default of 10% relative lift.

Quantify results are available from day 1

[5] – a Friday with 950k impressions



Publisher Name	Imps	vImps	Clicks	Conversions	vCTR %	CTC ‰	vVTC ‰
Flower	**15,509**	**8,684**	**18**	**0**	**0.20%**	**0.00‰**	**0.00‰**
Sloth	**372,413**	**157,524**	**105**	**17**	**0.06%**	**161.90‰**	**1.08‰**
MultiMeter	**533,901**	**180,816**	**479**	**81**	**0.24%**	**169.10‰**	**4.48‰**
RightButton	**32,122**	**9,685**	**27**	**0**	**0.27%**	**0.00‰**	**0.00‰**
	953,945	**356,709**	**629**	**98**	**0.16%**	**155.80‰**	**2.75‰**

Figure 2: Clearly, Sloth, one of the larger publishers, has a much worse vCTR (0.06%) than all the others (0.20% - 0.27%).

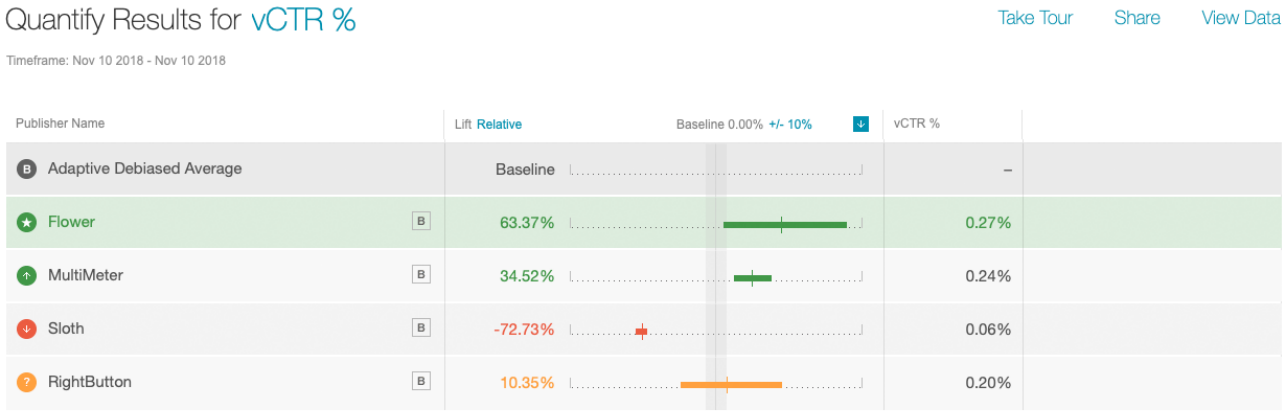


Figure 3: Results on day 1

We should collect more data. The actual credible intervals[6] based on the data are -35% to 65% for the challenger ***Right Button*** and 10% to 130% for ***Flower***[7]. In the following we suppress the numbers and just use the visual bars.

The visualisation stays similar as we collect more data, but the credible intervals become narrower. After **four days** it becomes pretty clear (Figure 4) that Flower, one of the smaller publishers, is actually the winner.

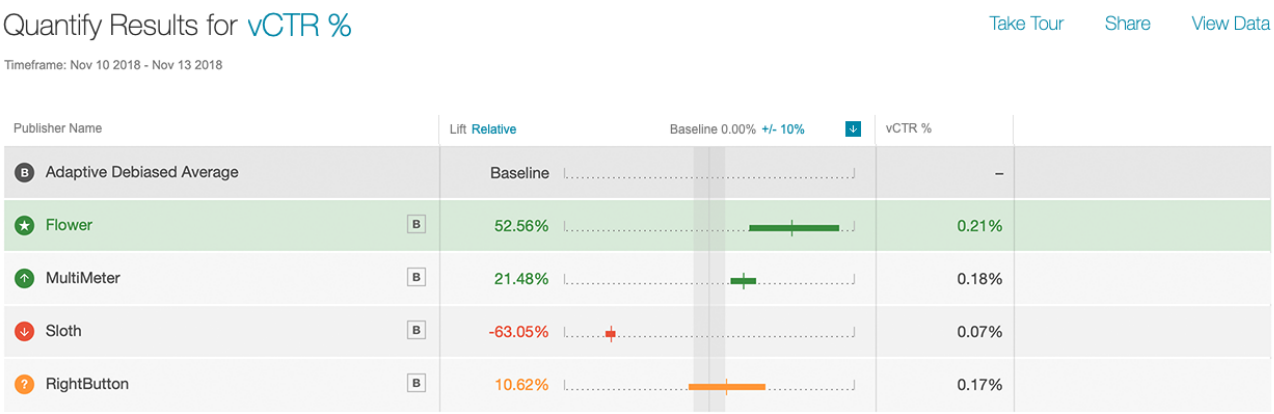


Figure 4: A clear winner

Speed

It took Quantify four days to detect a winner. Is that fast or slow? Let's compare with the classical significance test approach. In this classical setup, we have to pre-specify many parameters: the minimum detectable effect, a baseline conversion rate, and parameters alpha and beta controlling type I and II (false positive and false negative) errors.

Assuming we had guessed the baseline rate of 0.17%, looking for a minimum effect of 15% effect with typical values of 95% significance and 80% power, a sample size calculator tells us we should have 418,798 samples per variation.

In our example, after four days we had about 50k viewable impressions on the winning challenger, so we'd be looking at waiting for another twenty four days before we get results (at expected traffic 15k/day, since 420k/15k = 28 days in total). In more drastic terms, we come to an actionable conclusion in 4/28 = 14.2% of the time. That makes **Quantify 85,8%** faster than a significance test.

[6] – also known as probable ranges
[7] – users can hover over the row to expose these numbers

This comparison of course depends on the inputs we pick for the sample size calculator, varying estimated baseline and minimum effect, analogous calculations typically show **between 60% and 90% speedup**. Lower baseline or effect both increase the necessary sample size, and conversely.

> Quantify requires 60% to 90% less data

Equally important, in our opinion, is that a **classical significance test makes no statement on the actual lift measured in the test**, it simply decides *whether* there was a statistically significant level, given the assumptions made at the beginning of the test.

In general, we suggest to take a close look at credible intervals and to be conscious about influences like day-of-week or other effects before making **final** decisions. Figure 5 shows the results for our case after **nine days**, when all comparisons have come to a conclusion:

- our originally declared winner ***Flower*** remains the winner
- the early loser ***Sloth*** remains the loser
- the other challengers ***Rightlick*** and ***MultiMeter*** turn out to have similar performance as the debiased average

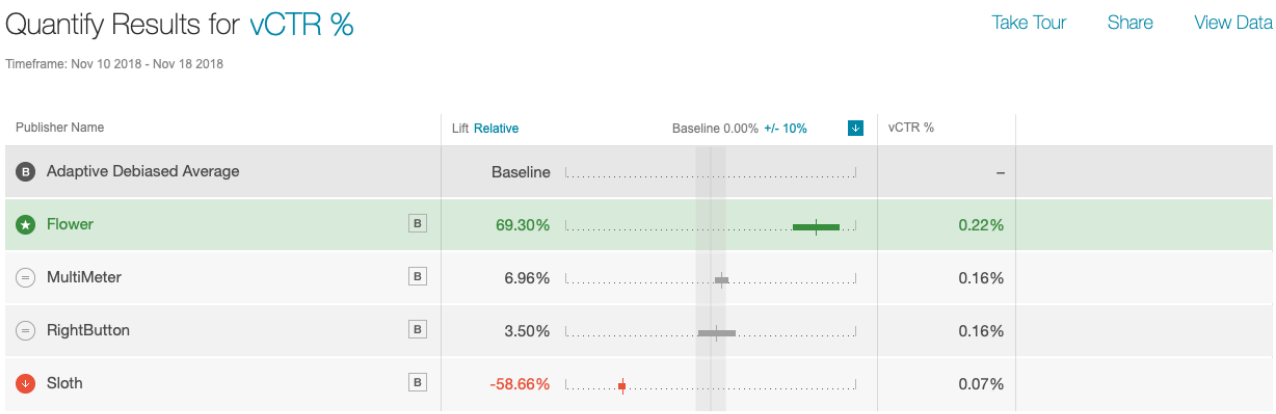


Figure 5: Credible intervals become narrower and can be clearly separated

Post-hoc data analysis

Another advantage of Quantify is that we can speculatively and post-hoc drill into performance metrics on any additional data points we have on the traffic (like contextual data, customer segments, device type, etc.). So the above experiment could also have been analyzed long after the data has been collected.

To illustrate the flexibility and interactivity we retrospectively "test" the vCTR lift of the challengers on creative level[8] over the baseline of all traffic (see Figure 6)

Quantify Results for vCTR %

Take Tour Share View Data

Timeframe: Nov 10 2018 - Nov 13 2018

Publisher Name	Creative Theme	Lift Relative	Baseline 0.00% +/- 10%	vCTR %
Adaptive Debiased Average	-	Baseline		–
MultiMeter B	Hippo Mobile	259.34%		0.63%
Flower B	Interactive	136.00%		0.45%
MultiMeter B	GoodPrice	25.96%		0.27%
Flower B	Hippo	-36.02%		0.14%
MultiMeter B	Dynamic	-53.47%		0.11%
RightButton B	GoodPrice	-55.90%		0.10%
MultiMeter B	Hippo	-57.39%		0.10%
Sloth B	Hippo	-62.60%		0.09%
Sloth B	Retargeted	-71.26%		0.07%
Flower B	GoodPrice Mobile	22.14%		0.25%

Figure 6: Post observational test broken down by additional dimension "Creative theme"

Some obvious take aways:

- the winning challenger Flower's performance is driven by one particular creative (Interactive)
- the publisher MultiMeter includes creatives of varying performance, it could be improved by allocating budget from the underperforming (Dynamic and Hippo) to the overperforming creative (GoodPrice), or by using the underperforming creatives' budget for new tests
- the creatives running on "Sloth" all perform bad, this publisher should be dropped unless there are other reasons for running with it, such as reach

[8] – Granularity: Publisher + Creative - this can be any granularity

3. Quantify vs. classical significance testing

It has been quipped that classical (frequentist) testing is for writing scientific articles, establishing the **existence of effects**. The main reason we think that this approach is inappropriate in advertising is that it **does not quantify the effect**.

> "Significance tests do not quantify the effect size"

Significance tests require elaborate preparation (usually supervised by a data scientist), such as specification of an appropriate significance level, power, effect to detect and null hypothesis. Even when things work out as expected, the outcome is only to "reject the null hypothesis of no effect". There is no estimate on the actual effect, which can lead to "statistically significant, practically insignificant" test results.

As any practitioner knows, there are further practical and theoretical issues:

1) since the goal is to prevent type I (false positive) and type II (false negative) errors, data must be used only once, otherwise the test assumptions are invalidated and must be repeated. Examples:

- testing for multiple variations (baseline data is used multiple times) requires careful corrections (e.g. Sidak correction)
- "peeking" at the test is strictly forbidden (data is used multiple times)
- post-hoc exploration is not possible (corresponds to potentially using the data infinitely many times - as many questions as one can come up with)
- incorrect test setup (optimistically setting too high of a minimum effect) leads to inconclusive tests, or tests that need to be repeated (at high cost)

2) the indirect nature of formulating a hypothesis with the aim of rejecting it, leads to many misunderstandings. Examples:

- incorrect assumption that 100% minus p-value is the probability that an effect exists
- "inconclusive" test does not imply there is no effect, just that it was not detected

3) lack of composability: whereas with Quantify, qualified and quantified "test" results can be composed (e.g., estimates of revenue lifts and spend differences can be combined into ROI estimates), significance tests for various effects cannot

4) often, required sample sizes / test run times are prohibitively long
One way of summarising this is to say that for applied purposes, classical significance testing focuses too much on type I and type II errors, while committing a grave type III error instead. Here, type III error is not a mathematical term, it just means: “obtaining the right answer to the wrong question”.

4. The math

Fundamentals

We take the Bayesian point of view that metrics are parameters (for which we have no direct observations) of likelihood models for underlying binary “success” processes, for which we ***do*** have observations[9]. These metrics themselves are random variables, with distributions that we should update according to all available data.

Further, we should

- explicitly formulate our full model, including prior assumptions (“beliefs”) about these metrics (in their interpretation as parameters)
- update our assumptions rationally according to the data we have access to, resulting in a posterior distribution over the parameters
- output summary statistics of the posterior distribution, containing at least a measure of centrality, and a measure of spread:
    - we pick the mode as centrality measure, due to its easy interpretation as “most likely value”
    - we pick the bounds of the 95% highest posterior density interval as our measure of spread, due to its property of being the smallest interval containing 95% of the probability density

In this whitepaper, we discuss only the simple (yet fundamental) case of the binary model: a simple Bernoulli (or binomial) model, with success rate parameter θ between 0 and 1. Using for robustness and simplicity Beta(1, 1) = U(0, 1) as uninformative (conjugate) prior quickly leads us to using the Beta distribution

$$p(\theta; \alpha + 1, \beta + 1) = \frac{\Gamma(\alpha+1,\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta+1)} \theta^{\alpha} (1 - \theta)^{\beta}$$

as our full posterior model, with α = number of successes, and β = number of failures observed in the data.

[9] – The approach extends to linear metrics such as time and money, cf. future white paper

As long as there is “enough” data, it overwhelms our choice of prior which, besides being uninformative, has the nice property that the posterior mode coincides with the “naive” quotient “successes/(successes + failures)”. In a low data situation, we might prefer to use an informative prior based on historical knowledge of success rates for a given metric.

Summing up, whenever we split our traffic into what we call **variations**, for each of the metrics in the next section we obtain the

- **variation mode**, the most likely value of the metric for the variation, given the data, and the
- **variation credible interval** (CI), the range of values containing the actual value of the metric with 95% probability

Lift calculation

For the lifts, we fundamentally make the **assumption** that the *variations are independent*. We can then calculate both the difference Y - X and normalised quotient (Y/X - 1) of the posterior distributions for each pair X, Y of variations, interpreted as standard beta random variables.

We call these the **absolute lift** and the **relative lift**, respectively, and associate

- **lift mode**, and
- **lift credible interval**

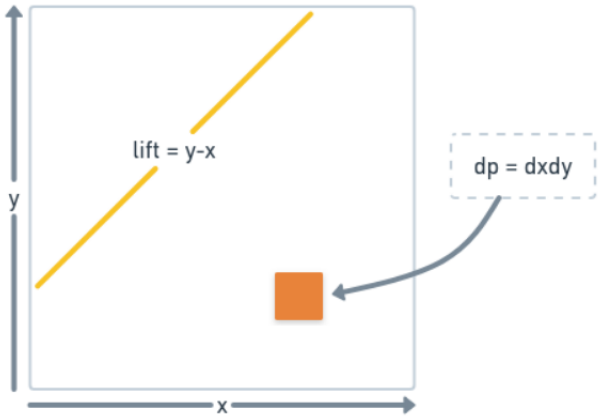
as centrality and spread measures of the lift.

For example, in the example of vCTR% from our Tour of Quantify, the relative lift is:

$$\frac{vCTR\%\ (challenger) - vCTR\%\ (baseline)}{vCTR\%\ (baseline)} = \frac{vCTR\%\ (challenger)}{vCTR\%\ (baseline)} - 100\%$$

Since there are $O(n^2)$ comparisons that can be made between n variations, we let the user pick one as baseline, and compare the others against it. This reduces the effort to O(n).

For evaluation, visually we can imagine the unit square. The point (x, y) corresponds to X having success rate x and Y having success rate y. Its probability is the product of probabilities of x and y being the success rates of X and Y. For given *l*, integrating this product density along the segment of points with $y=x+l$ gives us the probability density of the lift of Y over X being ***l***:



Theoretically, there are closed forms for both the beta difference (Pham-Gia et al. 1993, in terms of Appell's hypergeometric function of two variables) and beta quotient (Pham-Gia 2000, in terms of "the" hypergeometric function), for instance:

The random variable $W=X_1/X_2$ has density $D(\alpha_1,\alpha_2;\beta_1,\beta_2;w)$ given by

a) $f(w)=B(\alpha_1+\alpha_2, \beta_2)w^{\alpha_1-1}{}_2F_1(\alpha_1+\alpha_2,1-\beta_1;\alpha_1+\alpha_2+\beta_2;w) / A$, for $0<w\leq 1$

b) $f(w)=B(\alpha_1+\alpha_2, \beta_1)w^{-(1+\alpha_2)}{}_2F_1(\alpha_1+\alpha_2,1-\beta_2;\alpha_1+\alpha_2+\beta_1;1/w) / A$, for $w\geq 1$

using the hypergeometric[10] function ${}_2F_1$ and the constant $A = B(\alpha_1,\beta_1)B(\alpha_2,\beta_2)$.

Evaluating the modes and CIs then is just an automatic computation.

In practice, it is more efficient to perform Monte Carlo simulation, and just draw a bunch of samples.

[10] – https://en.wikipedia.org/wiki/Hypergeometric_function

Stopping rules

A stopping (or decision) rule is an "arbitrary" quantitative set of rules for making a decision, in particular it can be adjusted according to business priorities.

Given the lift metrics we just derived, we use a simple and "fast" approach: For a given "effect detection" parameter (typically, relative lift of challenger vs baseline) of, say 10%, we declare[11]:

- challenger wins: if entire credible interval for the lift is above 0%, and mode is above +10%
- baseline wins: if entire credible interval for the lift is below 0%, and mode is below -10%
- draw (conclude test): if entire credible interval for the lift is above -10% and below +10%
- continue test: in all other cases

Lift modeling vs interval comparison

It is more common to check "are the credible intervals disjoint" as a stopping rule. Our approach, using the modeled lift, has advantages:

- delivers faster results
- makes a qualified quantitative statement

Adaptive debiased averages as baseline

When calculating lifts, we need to have two variations: a baseline and a challenger. For initial exploration of the performance of a list of variations, it is useful to use some kind of "average" as baseline, and compare each variation against this average baseline.

The obvious choice is to use the "overall" performance of all variations taken together as this baseline. It has the "average" performance of all variations, weighted by their total number of observations. However, this approach has two drawbacks:

For each lift calculation, the **data** of the variation playing the role of challenger is **used twice**, once as input to the average, once as a challenger.

[11] – It would be interesting to do a frequentist analysis of significance and power levels of this procedure under appropriate assumptions.

- Typically, in a test there are one or two “incumbent” variations, and one or more challengers with lower traffic. By weighting according to number of observations, the “average” performance is biased towards the incumbent variation performances

- For this reason, we have developed and use an adaptive, debiased average as baseline. What is it? It is adaptive in the sense that for each variation we use a variation-specific average, depending only on the data of the other variations. Secondly, we set the mode of a challenger’s baseline to be the “naive” unweighted average of the other variations’ performance, with total number of observations kept at the actual total of the other variations’ observations.

Mathematically, it’s quite simple: If the variations are given by

$$V_i = (success_i,\ total_i), \quad \text{i} = 1..\text{N}$$

then for variation i we pick as baseline

$$B_i = ((\sum_{j \neq i} total_j) * (\sum_{j \neq i} success_j/total_j) / (N - 1)),\ \sum_{j \neq i} total_j)$$

In other words, we resample each “other” variation to have equal relative weight, and use the normal weighted average of these derived variations, as in the “obvious choice”.

Appendix A - Glossary

mode – most likely value of a distribution. Coincides with "num successes/total" in the binary case, under assumption of a flat prior.

credible interval – interval containing the true value of a distribution with a specified probability, by convention 95%

quantify – estimate a metric by listing both its mode and a credible interval

Appendix B - References

http://elem.com/~btilly/ab-testing-multiple-looks/part1-rigorous.html

http://varianceexplained.org/r/bayesian-ab-testing/

https://doingbayesiandataanalysis.blogspot.com/2013/11/optional-stopping-in-data-collection-p.html

Kruschke (2012), Bayesian testing supersedes the t-test,
http://www.indiana.edu/~kruschke/BEST/BEST.pdf

Gelman et al., BDA,
http://www.stat.columbia.edu/~gelman/book/

Gelman et al., ARM,
http://www.stat.columbia.edu/~gelman/arm/

Pham-Gia, Turkkan, Eng (1993): Bayesian analysis of the difference of two proportions,
https://doi.org/10.1080/03610929308831114

Pham-Gia (2000): Distributions of the ratios of independent beta variables and applications,
https://doi.org/10.1080/03610920008832632

admetrics GmbH
Hanauer Landstrasse 161-173
60314 Frankfurt/Main
Germany

www.admetrics.io