Sequential confidence intervals for relative lift with regression adjustments

Sven Schmit
Eppo
sven@geteppo.com

Evan Miller
Eppo
evan@geteppo.com

October 12, 2022

Abstract

In recent years more and more large companies have shifted to using more advanced techniques in their experimentation platform. Two aspects in particular stand out: variance reduction through techniques such as CUPED, and avoiding the peeking problem using sequential analysis. However, they are often treated independently, rather than as parts of a larger system. We describe our experimentation platform, which implements a combined solution to these problems: sequentially valid confidence intervals for relative lift, adjusted for pre-experiment data using a regression-based approach. This approach is able to handle a wide variety of metric types, empowers technical and non-technical users to make data-driven decisions without worrying about the peeking problem, and speeds up experiments by leveraging pre-experiment data.

1 Introduction

Due to advances in the modern data stack, collecting, processing, and analyzing data has become ever easier, allowing companies to leverage data earlier on in their life cycle. Combined with evidence of the value of an experimentation culture [Koning et al., 2019] we see that more and more companies are interested in running experiments to make data-driven decisions. Eppo is an experimentation platform with the aims of bringing state-of-the-art in-house experimentation tools, such as those built by Microsoft and Netflix, to the market for wide adoption. In order to be successful at this goal, such a platform has to be *general*; it cannot be focused on a particular use case, but has to cover a wide variety of situations, business models, metric types, and scales. Next, the platform should be *powerful* (in the statistical sense): often the amount of data we collect is a limiting factor, especially so for smaller companies, early stage companies, and B2B companies. Finally, the platform should be *intuitive*; to build an experimentation culture, we have to be able to provide insights that do not require a graduate degree in statistics to understand. In this work, we describe how we currently approach a core aspect of our platform, confidence intervals, which are the primary instrument to convey lift and uncertainty of experiment results on the platform. An example of the experiment results dashboard is shown in Figure 1.

For the reasons mentioned above, we provide sequential confidence intervals [Howard et al., 2020] on the relative lift so that experimenters can make decisions whenever they want, without having to abide by a series of caveats in order to avoid statistical traps (such as the peeking problem). To reduce variance we use regression adjustments Lin [2013] using a linear model with both pre-experiment metric data and dimensional data (features such as user persona, country etc.), While both sequential confidence intervals and the regression adjustment method are grounded in theory, in practice these assumptions can be hard to satisfy. Nonetheless, through simulations and A/A tests, we found that this is a powerful combination that works well in practice. There are a number of important aspects that we do not address here due to space

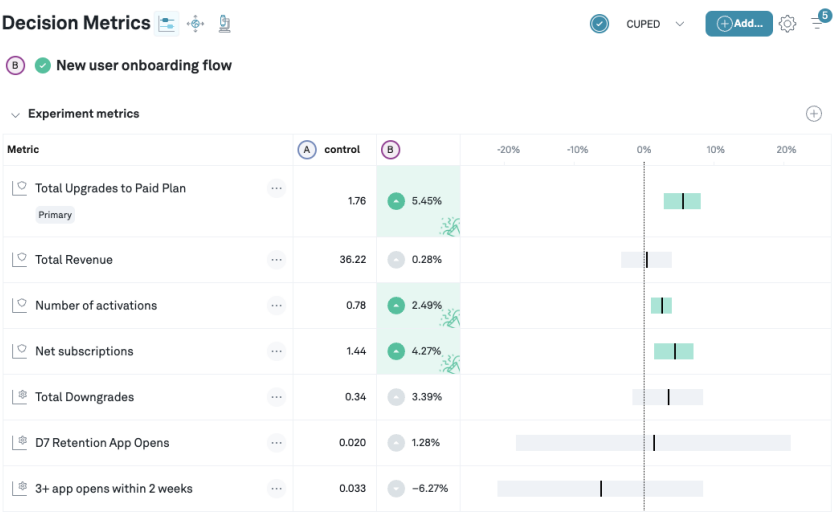

Figure 1: Dashboard showing results of a demo experiment in Eppo.

considerations, in particular: handling of outliers, non-stationarity of treatment effects (e.g. due to novelty bias or weekly seasonality) and multiple testing corrections.

2 Experiment setting

Let us take a slightly formal approach in defining the problem precisely, as it reveals subtle structure in how data is generated. Consider the standard setting of analyzing an experiment with two variations: treatment and control. During the experiment, subjects, indexed by $i = 1, \ldots, n$, are randomly assigned to treatment ($W_i = 1$) or control ($W_i = 0$). We are interested in understanding the differences between treatment and control on a set of metrics M, but focus on estimating the treatment effect for a single metric m. We observe events that correspond to metric m at specific timestamps for each subject throughout the experiment. For each subject, we can compute the metric value and subject i at time t as a function of all events (and timestamps) for that subject up until time t. For example, suppose we track spend events, then the revenue metric would sum spend amounts from the time of assignment up until time t. Alternatively, we can compute a 7-day conversion metric by looking at whether there is an event within 7 days of assignment.

We are now interested in understanding the difference in means for the underlying stochastic processes for treatment and control, with mean values $\mu_1(t)$ and $\mu_0(t)$. As metrics often differ in magnitude, we are interested in creating sequential confidence intervals around the relative lift

$$\tau(t) = \frac{\mu_1(t) - \mu_0(t)}{\mu_0(t)} = \frac{\mu_1(t)}{\mu_0(t)} - 1. \tag{1}$$

for the duration of the experiment. In practice, the results are updated periodically, e.g. once a day.

3 Sequential confidence intervals

Classical frequentist confidence intervals provide a type I error guarantee for a particular value of t. Sequential confidence intervals provide the much stronger guarantee of generating a *sequence* of confidence intervals with a type I error guarantee across the entire sequence [Howard et al., 2020]. Our approach to generating these confidence intervals is straightforward and flexible. First, we find estimators $\hat{\mu}_w(t)$ of $\mu_w(t)$ and derive

the asymptotic distribution. For now, let's simply use the sample averages for subjects with $W_i = w$. Below, we show how we can replace these simple sample averages with more powerful estimates using regression adjustments. In either case, we can combine the central limit theorem and delta method to find the asymptotic normal distribution of our effect estimate

$$\hat{\tau}(t) = \frac{\hat{\mu}_1(t)}{\hat{\mu}_0(t)} - 1 \to_D \mathcal{N}\left(\frac{\mu_1(t) - \mu_0(t)}{\mu_0(t)}, \sigma_\tau^2\right), \tag{2}$$

with

$$\sigma_\tau(t) = \frac{\mu_1(t)}{\mu_0(t)}\sqrt{\frac{\sigma_0^2(t)}{\mu_0^2(t)} + \frac{\sigma_1^2(t)}{\mu_1^2(t)}}. \tag{3}$$

From $\hat{\tau}$ it is straightforward to construct classical frequentist confidence intervals, but by default we produce sequential confidence intervals using the Normal Mixture bound in Howard et al. [2020] as

$$\hat{\tau}_l \pm \hat{\sigma}_\tau \sqrt{\frac{n+\rho}{n} \log\left(\frac{n+\rho}{\rho\alpha^2}\right)}, \tag{4}$$

$$\rho = \frac{M}{\log(\log(e\alpha^{-2})) - 2\log\alpha}. \tag{5}$$

Note that the assumptions required for sequential confidence intervals are quite strong in practice, but we found that the above confidence intervals perform well in both simulation studies as well as A/A tests. Here, M is a free parameter that can be used to tune for which sample size the confidence interval is relatively tightest, and hence it should be set approximately equal to the expected sample size for the experiment. α is the usual parameter to set the confidence level.

4 Variance reduction with regression adjustments

Many experimentation platforms have found great success in reducing variance and thus obtaining tighter confidence intervals by leveraging additional pre-experiment data, first popularized by Microsoft's CUPED implementation [Deng et al., 2013, Xie and Aurisset, 2016, Tang et al.]. CUPED computes the pre-experiment data on the metric of interest over a short window and uses that as a correction term to reduce variance. While powerful and simple to implement at scale, the standard CUPED approach has a couple of downsides: not all metrics have a straightforward pre-experiment equivalent; there are no benefits for experiments without pre-experiment data (such as sign-up flow experiments); handling missing values can cause issues; and pre-experiment data from other metrics may help reduce variance as well.

Instead, we follow the regression adjustment approach from Lin [2013], Wager [2020] to leverage for pre-experiment data. For each outcome metric (for which we want to estimate average treatment effect), we fit a regression on subjects in the control group, and similarly fit a regression on subjects in the treatment group. For each model, the dependent variable is the observed metric values for each subject in the experiment, and the independent variables are:

1. The pre-experiment metric values for the given outcome metric.
2. The pre-experiment metric values for *every other* metric in the experiment.
3. *Assignment dimensions*: Features such as country, browser, user persona, etc., which are available at the time the subject is first assigned to an experiment group

That is, given a metric m, we fit models for control and treatment for its value at time t based on

$$Y_{m,w}(t) = \beta_w X_w + \varepsilon \tag{6}$$

with

$$X = [1, X_{w,M}, X_{w,D}]. \tag{7}$$

$X_{w,M}$ captures pre-experiment metric data across a look-back period for all metrics in M, and $X_{w,D}$ contains pre-experiment assignment data. Let $p_{w,i}$ be the prediction of subject i from model w, which we can use to construct estimates $\hat{\mu}_w = \frac{1}{n}\sum_i p_{w,i}$. We can then leverage the following convergence result [Buja et al., 2019, Proposition 7]

$$\sqrt{n}(\hat{\mu}_w - \mu_w) \to_D \mathcal{N}(0, \sigma_w^2) \tag{8}$$

with

$$\sigma_w^2 = \mathbb{E}\left((Y_w - p_w(X))^2\right). \tag{9}$$

Note in particular that this result does not assume linearity of the population model, and furthermore that this approach can be extended to leverage modern machine learning methods. We can now use the estimate the mean square error based on our fitted values for control and treatment models as estimates of the variances, σ_w^2, plug them into Equation 2 and push this through the sequential confidence interval machinery.

Conveniently, when computing the average treatment effect we can leverage the linearity of expectation: $\mathbb{E}(Y_{i,1} - Y_{i,0}) = \mathbb{E}(Y_{i,1}) - \mathbb{E}(Y_{i,0})$, but this identity obviously does not hold for the relative lift: in general, $\mathbb{E}(Y_{i,1}/Y_{i,0}) \neq \mathbb{E}(Y_{i,1})/\mathbb{E}(Y_{i,0})$, In our case, we compute the latter, and one can use the Taylor expansion around $\mathbb{E}(Y_{i,0})$ to characterize the difference exactly.

While the regression model is computationally more demanding than CUPED, it is also more flexible and powerful. E.g. we can create indicator covariates for missing values, and leverage assignment dimensions, especially useful for experiments with new subjects (i.e. subjects for which we have no pre-experiment data). In practice, we add a small ridge regression penalty to be robust to multicollinearity issues. Furthermore, we fit the ridge regressions efficiently across many metrics by leveraging the fact that the covariate matrix X the same across all these models. This substantially speeds up computation, in particular when an experiment has many metrics configured.

5 Conclusion

We presented how we combine sequential confidence intervals and regression adjustments to generate confidence intervals. This is a powerful combination that can be leveraged across a wide variety of metric types and empower both technical and non-technical users to make informed decisions that are rooted in statistical rigor. The above method is currently running in production, powering results of experiments at both small (few metrics, thousands of subjects) and large (hundreds of metrics, millions of users) scales.

Acknowledgements

We would like to thank Steve Howard, Leo Pekelis, and Brian Karfunkel for their collaboration on both implementation of these methods as well as providing feedback on this work.

References

Andreas Buja, Richard A. Berk, Lawrence D. Brown, Edward I. George, Emily Pitkin, Mikhail Traskin, Linda H. Zhao, and Kai Zhang. Models as approximations i: Consequences illustrated with linear regression. *Statistical Science*, 2019.

Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *WSDM '13*, 2013.

Steven R. Howard, Aaditya Ramdas, Jon D. McAuliffe, and Jasjeet S. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 2020.

Rembrand Koning, Sharique Hasan, and Aaron K. Chatterji. Experimentation and startup performance: Evidence from a/b testing. *ERN: Firm Behavior & Competition (Topic)*, 2019.

W. Lin. Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. 2013.

Yixin Tang, Caixia Huang, David Kastelman, and Jared Bauman. Control using predictions as covariates in switchback experiments.

Stefan Wager. Lecture notes STATS 361: Casual inference, 2020.

Huizhi Xie and Juliette Aurisset. Improving the sensitivity of online controlled experiments: Case studies at netflix. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 645–654, 2016.