Percent Change Estimation in Large Scale Online Experiments

Jacopo Soriano

Google Inc.

November 3, 2017

Abstract

Online experiments are a fundamental component of the development of web-facing products. Given the large user-base, even small product improvements can have a large impact on an absolute scale. As a result, accurately estimating the relative impact of these changes is extremely important. I propose an approach based on an objective Bayesian model to improve the sensitivity of percent change estimation in A/B experiments. Leveraging pre-period information, this approach produces more robust and accurate point estimates and up to 50% tighter credible intervals than traditional methods. The R package abpackage provides an implementation of the approach.

1 Introduction

Tech companies like Amazon, Facebook, Google and Microsoft rely on A/B experiments to evaluate the impact of product changes. Examples of product changes range from new features in the user interface to algorithmic variations of the recommendation system.

In an A/B experiment a small subset of users is randomly assigned to the current state of the product (control group or group A) or to the potential change to the product (treatment group or group B). Once the experiment is completed, the experimenter is generally interested in testing whether the
average treatment effect is non-zero on key metrics like daily active users (DAUs), time on site, latency, revenue, etc. The most common and simplest statistical procedure for hypothesis testing in this context is the t-test. However, despite the large number of users in these experiments, power to detect a difference is often an issue when a 0.1% change can represent millions of dollars per year. As a result, in recent years variance reduction techniques ([Deng et al., 2013] and [Xie and Aurisset, 2016] and citations within) have been proposed to improve the statistical power of these tests.

Beyond hypothesis testing, it is important to estimate the size of the effect of the treatment. A good measure of the effect size is the percent change between the mean of the metric for the treatment group and the mean of the metric for the control group. Unlike the mean difference between the control and the treatment group, this quantity is scale free. Having a scale free parameter is attractive because it facilitates comparisons across different experiments and different metrics. Additionally, the percent change is interpretable. If one scales this experiment to 100% of traffic, for example, one expects an increase of x% in the metric of interest.

Percent change estimation has been largely discussed in the medical literature [Kaiser, 1989] and [Vickers, 2001], but it has not being carefully studied in the context of large scale online experiments. To my knowledge, [Kohavi et al., 2009] is the only work discussing percent change estimation, and variance reduction techniques are not considered. In the medical context the percent change is generally computed between the post-period and the pre-period in the same patient. In online experiments experimenters often observe large day-to-day variability in the metrics of interests. As a result, the percent change between the pre-period and the post-period is generally not meaningful, and instead the focus is on the percent change between treatment and control in the post-period. However, even if there is no direct interest in the pre-period, the pre-period can be incorporated into the analysis as a baseline covariate to improve inference in the post-period.

The main contribution of this paper is a sensitive statistical method for percent change estimation in A/B experiments that leverages the pre-period. Specifically, I propose a new approach based on an objective Bayesian model to estimate the percent change in the post-period while controlling for a single pre-period covariate. The resulting point estimates are substantially more precise, and the credible intervals (CIs) are up to 50% tighter than CIs based on traditional methods that do not correct for pre-period covariates.

Compared to classical methods, the advantages of using an objective
Bayesian approach are threefold. First, propagation of uncertainty across different stages of the model is very natural in a Bayesian set up. Secondly, inference on an arbitrary function of the parameters is also straightforward. In this case, the function is the percent change between the two means. Finally, by using objective priors, the experimenter has a statistical procedure with good frequentist properties without having to elicit prior parameters.

This paper is organized as follows. Section 2 provides some background and describes how to make inference on the percent change using an objective Bayesian approach, and common variance reduction techniques are briefly discussed. In Section 3 I extend the model to the case of a single baseline covariate, and I describe the algorithm for computing CIs. In Section 4 I illustrate the methodology on several real data examples. In Section 5 the main contributions of this work are summarized.

Code to use this methodology are freely available at https://google.github.io/abpackage in the form of an R package called abpackage.

2 Background

2.1 Difference and Percent Change

Assume that the observed values of the metric of interest for the treatment and the control groups are draws from the random variables $Y_{i,j}$ where $i = 1, \ldots, n_j$ and $j = c, t$. In large online experiments, data are often randomly bucketed, i.e., data are the result of an aggregation of the metric across many users [Chamandy et al., 2012]. For the DAUs metric, for example, the observation $Y_{i,t}$ represents the sum of the number of days that users that are in bucket $i$ and in the treatment group used the product across the $T$ days of the experiment. Since observations are the result of an aggregation across users and days, by the Central Limit Theorem it is reasonable to assume that the underlying random variables are normally distributed

$$Y_{i,j} \sim N(\mu_j, \sigma^2_j),$$

where $i = 1, \ldots, n_j$ and $j = c, t$. To test if there is a difference between the mean of the treatment group, $\mu_t$, and the mean of the control group, $\mu_c$, one can use a t-test.

In addition to identify whether there is an effect, often an experimenter is also interested in estimating the size of the effect. For strictly positive
metrics, a good measure of effect size is the percent change between the treatment mean and the control mean

\[ 100 \cdot \frac{\mu_t - \mu_c}{\mu_c} = 100 \cdot \frac{\mu_t}{\mu_c} - 100. \] (2)

As described earlier, this quantity is attractive because it is scale free, allowing natural comparison across several metrics within the same experiment or across experiments. The CI for (2) is well defined only when the estimate of the control mean \( \mu_c \) is far from zero. This is the case when the sample mean \( \bar{y}_c \) is sufficiently larger than the standard error \( SE_{\bar{y}_c} = \sqrt{s_c^2 / n_c} \), where \( s_c^2 \) is the sample variance of the control group. A simple rule of thumb that I would suggest is \( \bar{y}_c > 5 \cdot SE_{\bar{y}_c} \).

A simple way to estimate a complex function of parameters, such as (2), is to use a Bayesian approach. Given the posterior distribution for \( (\mu_c, \mu_t) \), one can sample posterior draws \( (\mu_c^{(d)}, \mu_t^{(d)}) \), compute \( 100 \cdot \frac{\mu_t^{(d)}}{\mu_c^{(d)}} - 100 \) for \( d = 1, \ldots, D \), and then compute the CI for (2) based on sample quantiles of such draws. The main limitations of this approach is that for each collection of draws the resulting CI will be slightly different due to Monte Carlo error. To avoid the variability introduced by sampling posterior draws, one can discretely approximate the posterior distribution on a fine grid, compute the percent change for each point on the grid, and then compute the CI based on the sample quantiles. This grid-based approach is recommended, and further details are provided on Section 3.2.

When using a Bayesian model, one must specify a prior for the unknown parameters. I consider the following two independent reference priors ([Bernardo, 1979]) for the control and treatment groups

\[ \pi(\mu_c, \sigma_c^2, \mu_t, \sigma_t^2) \propto \frac{1}{\sigma_c^2} \cdot \frac{1}{\sigma_t^2} \] (3)

Under this prior, the posterior is equal to

\[ (\mu_c, \mu_t | y_c, y_t) \sim T_{n_c-1}(\bar{y}_c, s_c) \times T_{n_t-1}(\bar{y}_t, s_t), \]

where:

- \( y_c = (y_{1,c}, \ldots, y_{n_c,c})' \) and \( y_t = (y_{1,t}, \ldots, y_{n_t,t})' \) indicate the observations of the two groups;

- \( \bar{y}_c \) and \( \bar{y}_t \) indicate the sample means of the two groups;
• \( s_c \) and \( s_t \) indicate the sample standard deviations of the two groups;

• \( T_x(y, z) \) indicates a non-standardized Student’s t-distribution with \( x \) degrees of freedom, location parameter \( y \) and scale parameter \( z \).

Priors (3) are also called matching priors because the resulting CIs for \( \mu_c \) and \( \mu_t \) match the frequentist confidence intervals for \( \mu_c \) and \( \mu_t \), respectively, for all significance levels \( \alpha \in (0, 1) \).

Alternative classical approaches include the Fieller’s method discussed in [Kohavi et al., 2009] and the delta method. The resulting CIs of these three methods are all very similar. The advantage of using a Bayesian approach is that the experimenter can replace the objective prior (3) with an informative prior when they have prior knowledge on the parameters of the model. In addition, the Bayesian model can be easily extended to include a baseline covariate. This extension is discussed in detail in Section 3.

### 2.2 Variance Reduction Techniques

Despite the large sample size of online experiments, power is an issue because the signal to noise ratio is often very small. Several variance reduction techniques have been proposed, with the most common being stratification and control-variates. See [Deng et al., 2013] and [Xie and Aurisset, 2016].

The idea of stratification is to separate users in several groups based on some categorical covariates. The covariates are chosen to have small variance within each stratum and high across strata. As [Najmi, 2016] pointed out, in practice this approach does not work because it is hard, if not impossible, to find such covariates.

A more successful technique is control-variate. The idea of control-variate is to identify a baseline metric \( X \) that has the same distribution for the treatment and the control group such that \( |\text{Cor}(X_{i,j}, Y_{i,j})| = |\rho_j| > 0 \) for \( j = t, c \).

Specifically, if \( (X_{i,j}, Y_{i,j}) \) is Normally distributed, and \( X_{i,j} \) has mean \( \mu_0 \) and variance \( \sigma_0^2 \), the conditional distribution has the following expression

\[
Y_{i,j} | X_{i,j} = x_{i,j} \sim N(\mu_j + \beta_j \cdot (x_{i,j} - \mu_0), \tau_j^2),
\]

where \( i = 1, \ldots, n_j, j = c, t, \beta_j := \rho_j \cdot \sigma_j / \sigma_0 \) and \( \tau_j^2 := (1 - \rho_j^2) \cdot \sigma_j^2 \).

Assume, for instance, that the correlation \( \rho_j > 0 \). Then, the expected mean for \( Y_{i,j} \) is higher/lower than the post-period mean \( \mu_j \) when \( x_{i,j} \) is
higher/lower than the baseline mean \( \mu_0 \). In other words, \( x_{i,j} \) is predictive of what one will observe in the post-period. As a result, the baseline metric \( X \) removes some of the variability of \( Y \). Specifically, the variance of \( Y \) is reduced by a factor \( 1 - \rho_j^2 \) with respect to the variance of \( \mu \). On one hand, the higher the correlation \( \rho_j \) is, the higher the variance reduction will be. On the other hand, when the two correlations are null, the two approaches have the same variance.

In general, a good metric \( X \) is the same metric as \( Y \) but computed prior to the start of the experiment. For example, for the DAUs metric, \( X_{i,j} \) would be the sum of the number of days that users in bucket \( i \) and in the condition group \( j \) used the product in the \( T' \) days preceding the start of the experiment. From now on, I will refer to \( X \) as the pre-period metric.

In online experiments it is not uncommon to observe correlations between the pre-period and the post-period of the order of 0.8, which result in width reductions of 40%. The length of the pre-period may have an impact on the correlation. In practice I observed that there are minimal incremental gains when considering a pre-period longer than a week, which is consistent with the findings in Deng et al., 2013.

3 Methodology

3.1 Model

In this section I describe the model to estimate the percent change between \( \mu_t \) and \( \mu_c \) when a pre-period metric \( X \) is available. The model has two components. The first component describes the observations in the pre-period.

\[
X_{i,j} \sim N(\mu_0, \sigma_0^2),
\]

where \( i = 1, \ldots, n_j \) and \( j = t, c \). The mean and the variance are identical for the treatment and the control since no treatment has occurred yet.

The second component describes the observations in the post-period given the pre-period

\[
Y_{i,j} | X_{i,j} = x_{i,j}, \mu_0 \sim N(\mu_j + \beta_j \cdot (x_{i,j} - \mu_0), \tau_j^2),
\]

where \( i = 1, \ldots, n_j \) and \( j = t, c \). Given \( x_j \) and \( \mu_0 \), \( Y_{i,j} \) follows a simple linear regression with slope \( \beta_j \), independent variable \( x_{i,j} - \mu_0 \) and residual error \( \epsilon_{i,j} \sim N(0, \tau_j^2) \), for each of the two groups \( j = t, c \).
The unknown parameters of the model have the following objective prior
\[ \pi(\mu_0, \sigma_0^2, \mu_t, \beta_t, \tau_t^2, \mu_c, \beta_c, \tau_c^2) \propto 1/\sigma_0^2 \cdot 1/\tau_t^2 \cdot 1/\tau_c^2. \]

The posterior distribution of \((\mu_t, \mu_c)\) can be easily computed using a Gibbs sampler. However, a Gibbs sampler can be computationally inefficient, particularly when one is interested in computing the posterior for a large number of metrics and experiments. To overcome this limitation, I developed a fast and deterministic algorithm which discretely approximates the posterior of \(\mu_t, \mu_c\) on a grid. A comparison of the proposed algorithm to a Gibbs sampler is provided in the Appendix.

3.2 Algorithm

In this section I describe the algorithm to approximate the posterior distribution of the model described in Section 3.1. The joint posterior distribution of \((\mu_t, \mu_c)\) can be written as follow
\[ \pi(\mu_t, \mu_c|x_t, x_c, y_t, y_c) = \int \left[ \prod_{j=t,c} \pi(\mu_j|x_j, y_j, \mu_0) \right] \pi(\mu_0|x_t, x_c, y_t, y_c) d\mu_0, \]
where \(x_j = (x_{1,j}, \ldots, x_{n_j,j})\) for \(j = t, c\).

There is no closed form for the exact posterior of \(\mu_0\), \(\pi(\mu_0|x_t, x_c, y_t, y_c)\).
However, since most of the information about \(\mu_0\) is contained in the pre-period, the exact posterior can be replaced by a pseudo-posterior which is computed only with respect to the pre-period data, i.e., \(\pi(\mu_0|x_t, x_c)\). Unlike the exact posterior, the distribution of the pseudo-posterior can be derived analytically.

A continuous univariate distribution can be discretely approximated by computing its quantiles. For example, the \(d\)th element of the discrete approximation will correspond to the quantile \((2 \cdot d - 1)/(2 \cdot D)\), and each element will have equal probability \(1/D\). By using such discretization, one can approximate the joint posterior distribution of \(\mu_0, \mu_c\) and \(\mu_t\). Altogether, this
gives us,

\[
\pi(\mu_t, \mu_c|x_t, x_c, y_t, y_c) \simeq \int \left[ \prod_{j=t,c} \pi(\mu_j|x_j, y_j, \mu_0) \right] \pi(\mu_0|x_t, x_c)d\mu_0
\]
\[
\simeq \frac{1}{D} \sum_{d_0=1}^{D} \prod_{j=t,c} \pi(\mu_j|x_j, y_j, \mu_0^{(d_0)})
\]
\[
\simeq \frac{1}{D^3} \sum_{d_0=1}^{D} \prod_{j=t,c} \sum_{d_j=1}^{D} \delta(\mu_j^{(d_j)}|x_j, y_j, \mu_0^{(d_0)}),
\]

where \(\delta(\cdot)\) represents the Dirac delta function, \(\mu_0^{(1)}, \ldots, \mu_0^{(D)}\) represent the discretization of the pseudo-posterior distribution of \(\mu_0\), and \((\mu_j^{(1)}|x_j, y_j, \mu_0^{(d_0)}), \ldots, (\mu_j^{(D)}|x_j, y_j, \mu_0^{(d_0)})\) indicate the discretization of the conditional distributions of \(\mu_j\) given \(\mu_0 = \mu_0^{(d_0)}\) for \(j = c, t\).

In the next two paragraphs the posterior distributions of the pre-period mean \(\pi(\mu_0|x_t, x_c)\) and post-period means \(\pi(\mu_j|x_j, y_j, \mu_0)\) for \(j = t, c\) are discussed.

**Pre-period mean estimation** Under the matching prior \(\pi(\mu_0, \sigma_0^2) \propto 1/\sigma_0^2\), the distribution of the pseudo-posterior of \(\mu_0\) is equal to

\[
T_{n_c+n_t-1}(\bar{y}_0, s_0),
\]

where \(\bar{y}_0\) represents the sample mean of the two groups combined in the pre-period, and \(s_0\) indicates the sample standard deviation of the two groups combined in the pre-period.

**Experiment period means estimation** As discussed earlier, given \(x_j\) and \(\mu_0, Y_{i,j}\) follows a simple linear regression with independent variable \(x_{i,j} - \mu_0\) and residual variance \(\tau_j^2\), for each of the two groups \(j = t, c\).

Since the priors for \((\mu_c, \beta_c, \tau_c^2)\) and \((\mu_t, \beta_t, \tau_t^2)\) are mutually independent, than, given \(\mu_0\), the posterior distributions of \(\mu_c\) and \(\mu_t\) are also mutually independent. Specifically, given \(\mu_0\), the posterior for \(\mu_j\) is equal to

\[
\mu_j|x_j, y_j, \mu_0 \sim T_{n_j-2}(\hat{\mu}_j, \hat{\tau}_j \cdot z_j)
\]
\[
z_j := \left( \sum_{i=1}^{n} x_{i,j}^2 \right) / \left( n_j \cdot (n_j - 1) \cdot s_{0,j}^2 \right),
\]

8
where $\hat{\mu}_j$ is the ordinary least squares estimate of $\mu_j$, $\hat{\tau}_j$ is the ordinary least squares estimate of the standard deviation of the residuals, and $s^2_{0,j}$ is the sample variance in the pre-period for group $j$, where $j = c, t$. Note that $\hat{\mu}_j = \hat{\mu}_j(x_j, y_j, \mu_0)$ and $\hat{\tau}_j = \hat{\tau}_j(x_j, y_j, \mu_0)$, since they are estimates of a simple linear regression with dependent variable $y_{i,j}$ and independent variable equal to $x_{i,j} - \mu_0$.

Similar to the pre-period mean, the distributions of the means of the control group and the treatment group are discretized based on their quantiles. Given a discrete approximation of the pre-period mean, the two distributions are conditionally independent, and their joint distributions can be approximated by the cross product of the two discrete approximations.

Figure 1 shows how the discrete approximation of $\pi(\mu_t, \mu_c|x_t, x_c, y_t, y_c)$ is constructed through a toy example based on a simulated dataset. The scatterplot on the left shows the pairs $(\mu_t^{(d)}, \mu_c^{(d)})$ for $d = 1, \ldots, D^3 = 20^3$, and the histogram on the right shows the percent change for each of them, i.e., $100 \cdot \mu_t^{(d)} / \mu_c^{(d)} - 100$, and estimate of the 90% CI based on the sample quantiles.

The discrete approximation of $\pi(\mu_t, \mu_c|x_t, x_c, y_t, y_c)$ provides robust estimates of the lower and upper extremes of a CI for $D \approx 50$. An analysis of the robustness of the estimates as a function of the number of nodes $D$ is provided in the Appendix.

## 4 Empirical Results

In this section I assess the performance of the methodology using two examples. In each example I compare the results using only post-period data (Post) to the results with a pre-period correction applied (Pre-Post). In the first example I study the robustness of the methodology as a function of the level of misalignment between the control and treatment group in the pre-period. In the second example I study the width of the CIs as a function of the traffic size and the number of days of the experiment. Both examples are based on YouTube experiments and the DAUs metric. Similar results are obtained for other metrics and are for this reason omitted. The two approaches provide comparable results for metrics capturing user behaviors that are not consistent over time.
Figure 1: In the scatterplot on the left each dot indicates a pair \((\mu_t(d), \mu_c(d))\) for \(d = 1, \ldots, D^3 = 20^3\). The dots are colored based on the 20 values \(\mu_0^{(1)}, \ldots, \mu_0^{(20)}\). The histogram on right indicates the approximation to the posterior distribution for the percent change between \(\mu_t\) and \(\mu_c\) based on the 203 pairs \((\mu_t^{(d)}, \mu_c^{(d)})\). Each of the 100 bins has equal probability. The 90 central bins are colored in grey to visually identify, as an example, the 90\% CI.

4.1 Robustness to Pre-Period Misalignment

This example is based on 100,000 A/A YouTube experiments. In A/A experiments the true percent change is known, and so it is possible to assess both the coverage of CIs and the mean squared error (MSE) of the point estimates.

I compute the permutation p-value for the mean difference in the pre-period, and then I bucket all the experiments in 100 groups based on their pre-period p-value. The first group contains all experiments where the pre-period p-value \(\in [0, 0.01]\), the second groups contains all experiments where the pre-period p-value \(\in [0.01, 0.02]\), etc. In the pre-period there is no treatment applied to the treatment group, and so any difference in the distribution of the two groups is due to chance. Thus, the pre-period p-value has a uniform distribution, resulting in approximately 1,000 experiments per pre-period p-value bucket.
For each pre-period $p$-value bucket Figure 2 shows the empirical coverage of the 95% CI and the MSE in the post period. Pre-Post outperforms Post in both coverage robustness and MSE. Both Pre-Post and Post have the right coverage on average. However, the coverage of Pre-Post is uniform with respect to the pre-period $p$-value, while the coverage of Post is strongly dependent on the pre-period permutation $p$-value. In particular, when the pre-period $p$-value is smaller than 0.1, i.e., the two groups are not well aligned in the pre-period, Post has substantially lower coverage than the nominal coverage. Similarly, the MSE of Pre-Post is uniform, while the MSE of Post is very high when the pre-period $p$-value is small and comparable to Pre-Post when the $p$-value is large.

Figure 2: On the left, coverage of the CIs in the post-period as a function of the permutation $p$-value for the pre-period. On the right, MSE for the post-period as a function of the permutation $p$-value for the pre-period. The red line indicates Pre-Post, and the blue line indicates Post. The dashed black line in the left plot indicates 95%, the nominal coverage of the CIs.

4.2 Width of the Credible Intervals

In this example I study the impact of traffic size and number of days of experiment on the DAUs metric. This example is based on 40 identical YouTube experiments, where 20 experiments have twice the traffic ($2p$) of
the other 20 experiments \((p)\). The pre-period corresponds to the 7 days prior to the start of the experiment.

In Figure 3 each thin line represents an individual experiment, and the thick line represents the average across experiments. Each color represents a different method and traffic size combination. Specifically, red corresponds to Post based on traffic proportion \(p\), green to Post based on twice the traffic \((2p)\), blue to Pre-Post based on traffic proportion \(p\), and purple to Pre-Post based on twice the traffic \((2p)\).

Correcting for the pre-period results in substantially tighter CIs. In this example the CIs for Pre-Post with traffic proportion \(p\) are tighter than the CI of Post with traffic proportion \(2p\). In addition, even if one takes into account the week spent to run the pre-period, the CIs of Pre-Post are still substantially tighter than the CIs of Post. Specifically, the CIs of Pre-Post after 1 day of the experiment are tighter than the CIs of Post based on 8 or more days of the experiment.

The rate at which the width of the CIs decays is slower than the inverse of the square root of the number of days. This is due to the fact that metrics are generally not independent across days, but instead are positively correlated over time. This implies that unless the treatment effect increases over time, the gains of running a longer experiments are smaller than those of running a larger experiment. Given the fact that a large number of experiments are often running in parallel (Tang et al., 2010), it is not possible to have a large sample size for each of them. Thus, maximizing the sensitivity of each individual experiment is crucial.

5 Conclusion

Pre-Post can substantially increase the sensitivity of large scale online experiments relative to existing approaches. The width of CIs associated with YouTube experiments, for example, can be reduced by up to 50%. This creates the opportunity for faster experimental cycles through shorter experiments, or exposing a smaller fraction of users to the experiment, while maintaining the same statistical power. Similarly, one can substantially increase the power to detect small effects, while maintaining the same length of the experiment and the same traffic size. Detecting small effects is important for web-facing products where the sum of several small improvements on a relative scale may have an overall large impact on an absolute scale.
Figure 3: Width of CIs as a function of the number of days of the experiment. Red corresponds to Post based on traffic proportion $p$, green corresponds to Post based on traffic proportion $2p$, blue corresponds to Pre-Post based on traffic proportion $p$, and purple corresponds to Pre-Post based on traffic proportion $2p$. The thin lines indicate individual experiments, and the thick lines indicate the average across experiments.

The main limitation of Pre-Post is the necessity to run a pre-period for a week before the start of the experiment. To overcome this limitation, YouTube developed an experiment framework to retrospectively compute pre-periods. Under this framework, for all users exposed to the treatment or the control group in the post-period, the pre-period metric $X$ is automatically and retrospectively computed. As a result, the experimenter does not need to set up a pre-period and can immediately start the experiment, resulting in simpler and faster experimental cycles.

Pre-period metrics can also be used to constantly monitor the health of the experiment system. Systematic deviations in the pre-period metrics between the treatment and the control can be an indication, for example, of
issues in the traffic diversion or the logging.

An open source R package implementing the methods described in the paper is freely available at [https://google.github.io/abpackage](https://google.github.io/abpackage).

**References**

[Bernardo, 1979] Bernardo, J. M. (1979). Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 113–147.

[Chamandy et al., 2012] Chamandy, N., Muralidharan, O., Najmi, A., and Naidu, S. (2012). Estimating uncertainty for massive data streams. *Google Internal technical report*.

[Deng et al., 2013] Deng, A., Xu, Y., Kohavi, R., and Walker, T. (2013). 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.

[Kaiser, 1989] Kaiser, L. (1989). Adjusting for baseline: change or percentage change? *Statistics in medicine*, 8(10):1183–1190.

[Kohavi et al., 2009] Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18(1):140–181.

[Najmi, 2016] Najmi, A. (2016). LSOS experiments: how I learned to stop worrying and love the variability. *The Unofficial Google Data Science Blog*.

[Tang et al., 2010] Tang, D., Agarwal, A., O’Brien, D., and Meyer, M. (2010). Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 17–26. ACM.

[Vickers, 2001] Vickers, A. J. (2001). The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. *BMC medical research methodology*, 1(1):6.
Appendix

Comparison of the Deterministic Algorithm and a Gibbs Sampler

In this section I compare the deterministic algorithm (DA) presented in Section 3.2 and a Gibbs sampler (GS) using two examples.

The first example is a case study to understand the robustness of the estimates as a function of the number of nodes $D$ for the DA and the number of iterations for the GS. In the second example I compare the point estimates, the CI widths and the computing times of the DA and the GS over 100 simulated datasets.

For a single simulated dataset, the plot on the left of Figure 4 shows the estimates of the 97.5%, the 50.0% and the 2.5% quantiles of the percent change (2) as a function of the number of nodes $D$. The estimate of the 50% quantile is immediately stable, while the estimates of the other two quantiles are stable starting from $D \approx 50$. The plot on the right shows the estimates of the same quantiles based on three independent Markov chains, where different colors are used to identify the different Markov chains. The chains are computed using a GS with a burn-in of 100 iterations and no thinning. Once the Markov chains have accumulated $\approx 2000$ iterations, the estimates of the quantiles are stable.

In the second example, the number of iterations of the GS is fixed to 2000 and the number of nodes of the DA to 50, and I compare the estimates and computing times of the DA and the GS over 100 datasets. The datasets are simulated from the following model

$$X_{i,j} \sim N(100, 1)$$
$$Y_{i,j} \sim N(100 + 10I_{(j=t)}, 1),$$

where $\text{cor}(X_{i,j}, Y_{i,j}) = 0.8$, $i = 1, \ldots, 20$ and $j = c, t$. The results of the analysis are provided in Table 1. The first two rows show the average CI
Figure 4: For a simulated dataset, the plot on the left shows the estimates of 97.5%, the 50.0% and the 2.5% quantiles of the percent change in function of the number of nodes $D$. The plot on the right shows the same estimates based on three Markov chains, where each color indicates a different Markov chain.

The main difference between the two approaches is the computing time with the DA outperforming the GS. The average computing time for the DA is 0.16 seconds, while the average computing time for the GS is 36.56 seconds. The code was written in R for both approaches, and it was tested on an Intel Xeon CPU E5-1650 v3 at 3.50GHz.
Table 1: The first two rows show the average CI width, point estimate and computing time for the DA and the GS across the 100 datasets, and the third row shows the average differences between DA and GS. The numbers in parenthesis represent the standard deviations.

| method | CI width, % | point estimate, % | time, sec |
|--------|-------------|-------------------|-----------|
| DA     | 0.84 (0.10) | 10.02 (0.22)      | 0.16 (0.02) |
| GS     | 0.86 (0.10) | 10.02 (0.22)      | 36.56 (0.23) |
| DA-GS  | -0.02 (0.02)| 0.00 (0.01)       | -36.41 (0.23) |