Applying the Delta Method in Metric Analytics: A Practical Guide with Novel Ideas

Alex Deng
Microsoft Corporation
Redmond, WA
alexdeng@microsoft.com

Ulf Knoblich
Microsoft Corporation
Redmond, WA
ulfk@microsoft.com

Jiannan Lu*
Microsoft Corporation
Redmond, WA
jiannl@microsoft.com

ABSTRACT

During the last decade, the information technology industry has adopted a data-driven culture, relying on online metrics to measure and monitor business performance. Under the setting of big data, the majority of such metrics approximately follow normal distributions, opening up potential opportunities to model them directly without extra model assumptions and solve big data problems via closed-form formulas using distributed algorithms at a fraction of the cost of simulation-based procedures like bootstrap. However, certain attributes of the metrics, such as their corresponding data generating processes and aggregation levels, pose numerous challenges for constructing trustworthy estimation and inference procedures. Motivated by four real-life examples in metric development and analytics for large-scale A/B testing, we provide a practical guide to applying the Delta method, one of the most important tools from the classic statistics literature, to address the aforementioned challenges. We emphasize the central role of the Delta method in metric analytics by highlighting both its classic and novel applications.

KEYWORDS

A/B testing, big data, distributed algorithm, large sample theory, online metrics, randomization, quantile inference, longitudinal study

ACM Reference Format:
Alex Deng, Ulf Knoblich, and Jiannan Lu. 2018. Applying the Delta Method in Metric Analytics: A Practical Guide with Novel Ideas. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3219919

1 INTRODUCTION

1.1 Background

The era of big data brings both blessings and curses [21]. On one hand, it allows us to measure more accurately and efficiently, to study smaller and more subtle effects, and to tackle problems with smaller signal-to-noise-ratio. On the other hand, it demands larger storage and more intensive computations, forcing the data science community to strive for efficient algorithms which can run in parallel in a distributed system. Many existing algorithms and methods (e.g., support vector machines) that are known to work well in small data scenarios do not scale well in a big data setting [12, 25, 47]. Recently, there has been an increasing amount of research interest in meta-algorithms, which can extend algorithms that are difficult to parallelize into distributed algorithms [6, 52], and ideas that resemble the divide-and-conquer algorithm [31, 34].

At the same time, there is a class of algorithms which are trivially parallelizable and therefore can downsize big data problems into smaller ones. The key idea behind them, which dates back to Fisher [24], is to summarize the original data set by a low-dimensional vector of summary statistics, which can often be computed in a parallel and distributed way. For example, to estimate the mean and variance of a normal distribution from independent and identically distributed (i.i.d.) observations, we only need to obtain their sum and sum of squares, which are the corresponding summary statistics[1] and can be trivially computed in a distributed fashion. In data-driven businesses such as information technology, these summary statistics are often referred to as *metrics*, and used for measuring and monitoring key performance indicators [15, 18]. In practice, it is often the changes or differences between metrics, rather than measurements at the most granular level, that are of greater interest. In the context of randomized controlled experimentation (or A/B testing), inferring the changes of metrics can establish causality [36, 44, 50], e.g., whether a new user interface design causes longer view times and more clicks.

1.2 Central limit theorem and Delta method

We advocate directly modeling the metrics rather than the original data-set. When analyzing data at the most granular level (e.g., user), we need some basic assumptions of the underlying probabilistic model, such as i.i.d. observations. When looking at the metrics level, we also need to know their joint distribution. This is where the blessing of big data comes into play. Given large sample sizes, the metrics often possess desirable asymptotic properties due to the *central limit theorem* [45]. To ensure that the paper is self-contained, we first review the central limit theorem in its most well-known form. Let $X_1, \ldots, X_n$ be n i.i.d. observations with finite mean μ and variance $\sigma^2 > 0$. We let $\bar{X}$ denote the sample average, then as the sample size $n \to \infty$, in distribution

$$\sqrt{n}(\bar{X} - \mu)/\sigma \to N(0, 1).$$

*The authors are listed in alphabetical order.

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3219919

[1] In fact, they are *sufficient statistics* [10], i.e., they can represent the original data-set perfectly without losing any information.

A common application of the central limit theorem is to construct the $100(1-\alpha)\%$ confidence interval of μ as $\bar{X} \pm z_{\alpha/2}\sigma/\sqrt{n}$, where $z_{\alpha/2}$ is the $(\alpha/2)$th quantile for $N(0,1)$. This is arguably one of the most influential results of asymptotic statistics used in almost all scientific fields whenever an estimate with error bars is presented.

While an influential result, the central limit theorem in its basic form only applies to the average of i.i.d. random variables, and in practice our metrics are often more complex. To enjoy the blessing of big data, we employ the Delta method, which extends the normal approximations from the central limit theorem broadly. For illustration we only review the uni-variate case. For any random variable T_n (the subscript indicates its dependency on n, e.g., sample average) and constant θ such that $\sqrt{n}(T_n - \theta) \to N(0,1)$ in distribution as $n \to \infty$, the Delta method allows us to extend its asymptotic normality to any continuous transformation $\phi(T_n)$. To be more specific, by using the fact that $T_n - \theta = O(1/\sqrt{n})$ and the first order Taylor expansion [43]

$$\phi(T_n) - \phi(\theta) = \phi'(\theta)(T_n - \theta) + O\{(T_n - \theta)^2\}, \tag{1}$$

we have in distribution

$$\sqrt{n}\left\{\phi(T_n) - \phi(\theta)\right\} \to N\left\{0, \phi'(\theta)^2\right\}.$$

This is the Delta method. Without relying on any assumptions other than "big data," the Delta method is *general*. It is also *memorable* – anyone with basic knowledge of calculus can derive it. Moreover, the calculation is trivially *parallelizable* and can be easily implemented in a distributed system. Nevertheless, although conceptually and theoretically straightforward, the *practical* difficulty is to find the right "link" function ϕ that transforms the simple average to our desired metric. Because of different attributes of the metrics, such as the underlying data generating process and aggregation levels, the process of discovering the corresponding transformation can be challenging. However, unfortunately, although various applications of the Delta method have previously appeared in the data mining literature [16, 36, 39, 44], the method itself and the discovery of ϕ were often deemed technical details and only briefly mentioned or relegated to appendices. Motivated by this gap, we aim to provide a practical guide that highlights the effectiveness and importance of the Delta method, hoping to help fellow data scientists, developers and applied researchers conduct trustworthy metric analytics.

1.3 Scope and contributions

As a practical guide, this paper presents four applications of the Delta method in real-life scenarios, all of which have been deployed in Microsoft's online A/B testing platform ExP [35, 36] and employed to analyze experimental results on a daily basis:

- In Section 2, we derive the asymptotic distributions of ratio metrics[2]. Compared with standard approaches by Fieller [22, 23], the Delta method provides a much simpler and yet almost equally accurate and effective solution;
- In Section 3, we analyze cluster randomized experiments, where the Delta method offers an efficient alternative algorithm to standard statistical machinery known as the mixed effect model [4, 26], and provides unbiased estimates;
- In Section 4, by extending the Delta method to outer confidence intervals [42], we propose a *novel* hybrid method to construct confidence intervals for quantile metrics with almost no extra computational cost[3]. Unlike most existing methods, our proposal does not require repeated simulations as in bootstrap, nor does it require estimating an unknown density function, which itself is often a theoretically challenging and computationally intensive task, and has been a center of criticism [7];
- In Section 5, we handle missing data in within-subject studies by combining the Delta method with data augmentation. We demonstrate the effectiveness of "big data small problem" approach when directly modeling metrics. Comparing to alternative methods that need to put up a model for individual subjects, our method requires less model assumptions.

The main purpose of this paper is to promote the Delta method as a *general*, *memorable* and *efficient* tool for metric analytics. In particular, our contributions to the existing data mining literature include:

- Practical guidance for scenarios such as inferring ratio metrics, where the Delta method offers scalable and easier-to-implement solutions;
- A novel and computationally efficient solution to estimate the sampling variance of quantile metrics;
- A novel data augmentation technique that employs the Delta method to model metrics resulting from within-subject or longitudinal studies with unspecified missing data patterns.

For reproducibility and knowledge transfer, we provide all relevant computer programs at https://aka.ms/exp/deltamethod.

2 INFERRING PERCENT CHANGES

2.1 Percent change and Fieller interval

Measuring *change* is a common theme in applied data science. In online A/B testing [15, 17, 36, 37, 44], we estimate the average treatment effect (ATE) by the difference of the same metric measured from treatment and control groups, respectively. In time series analyses and longitudinal studies, we often track a metric over time and monitor changes between different time points. For illustration, let $X_1, \ldots, X_n$ be i.i.d. observations from the control group with mean μ_x and variance σ_x^2, $Y_1, \ldots, Y_n$ i.i.d. observations from the treatment group with mean μ_y and variance σ_y^2, and σ_{xy} the covariance[4] between X_i's and Y_j's. Let $\bar{X} = \sum_{i=1}^n X_i/n$ and $\bar{Y} = \sum_{i=1}^n Y_i/n$ be two measurements of the same metric from the treatment and control groups, respectively, and their difference $\hat{\Delta} = \bar{Y} - \bar{X}$ is an unbiased estimate of the ATE $\Delta = \mu_y - \mu_x$. Because both $\bar{X}$ and $\bar{Y}$ are approximately normally distributed, their difference $\hat{\Delta}$ also follows an approximate normal distribution with mean Δ and variance

$$\mathrm{Var}(\hat{\Delta}) = (\sigma_y^2 + \sigma_x^2 - 2\sigma_{xy})/n.$$

[2] Pedantically, ratio of two measurements of the same metric, from the treatment and control groups.

[3] Our procedure computes two more quantiles for confidence interval. Since the main cost of quantile computing is sorting the data, computing three and one quantiles cost almost the same.

[4] For A/B testing where the treatment and control groups are independently sampled from a super population, $\sigma_{xy} = 0$.

Consequently, the well-known 100(1-α)% confidence interval of Δ is $\hat{\Delta} \pm z_\alpha \times \widehat{\text{Var}}(\hat{\Delta})$, where $\widehat{\text{Var}}(\hat{\Delta})$ is the finite-sample analogue of $\text{Var}(\hat{\Delta})$, and can be computed using the sample variances and covariance of the treatment and control observations, denoted as s_y^2, s_x^2 and s_{xy} respectively.

In practice, however, *absolute* differences as defined above are often hard to interpret because they are not scale-invariant. Instead, we focus on the *relative* difference or percent change $\Delta\% = (\mu_y - \mu_x)/\mu_x$, estimated by $\hat{\Delta}\% = (\bar{Y} - \bar{X})/\bar{X}$. The key problem of this section is constructing a 100(1-α)% confidence interval for $\hat{\Delta}\%$. For this classic problem, Fieller [22, 23] seems to be the first to provide a solution. To be specific, let $t_r(\alpha)$ denote the $(1-\alpha)$th quantile for the $t-$distribution with degree of freedom r, and $g = ns_x^2 t_{\alpha/2}^2(r)/\bar{X}^2$, then Fieller's interval of $\Delta\%$ is

$$\frac{1}{1-g}\left\{\frac{\bar{Y}}{\bar{X}} - 1 - \frac{g s_{xy}}{s_x^2} \pm \frac{t_{\alpha/2}(r)}{\sqrt{n}\bar{X}}\sqrt{s_y^2 - 2\frac{\bar{Y}}{\bar{X}}s_{xy} + \frac{\bar{Y}^2}{\bar{X}^2}s_x^2 - g\left(s_y^2 - \frac{s_{xy}^2}{s_x^2}\right)}\right\} \tag{2}$$

Although widely considered as the standard approach for estimating variances of percent changes [38], deriving (2) is cumbersome (see [46] for a modern revisit). The greater problem is that, even when given this formula, applying it often requires quite some effort. According to (2) we need to not only estimate the sample variances and covariance, but also the parameter g.

2.2 Delta method and Edgeworth correction

The Delta method provides a more intuitive alternative solution. Although they can be found in classic textbooks such as [10], this paper (as a practical guide) still provides all the relevant technical details. We let $T_n = (\bar{Y}, \bar{X})$, $\theta = (\mu_y, \mu_x)$ and $\phi(x, y) = y/x$. A multivariate analogue of (1) suggests that $\phi(T_n) - \phi(\theta) \approx \nabla\phi(\theta)\cdot(T_n - \theta)$, which implies that

$$\frac{Y}{X} - \frac{\mu_y}{\mu_x} \approx \frac{1}{\mu_x}(\bar{Y} - \mu_y) - \frac{\mu_y}{\mu_x^2}(\bar{X} - \mu_x) \tag{3}$$

For $i = 1, \ldots, n$, let $W_i = Y_i/\mu_x - \mu_y X_i/\mu_x^2$, which are also i.i.d. observations. Consequently, we can re-write (3) as $\bar{Y}/\bar{X} - \mu_y/\mu_x \approx \sum_{i=1}^n W_i/n$, leading to the Delta method based confidence interval

$$\underbrace{\frac{\bar{Y}}{\bar{X}} - 1}_{\text{point estimate}} \quad \underbrace{\pm \frac{z_{\alpha/2}}{\sqrt{n}\bar{X}}\sqrt{s_y^2 - 2\frac{\bar{Y}}{\bar{X}}s_{xy} + \frac{\bar{Y}^2}{\bar{X}^2}s_x^2}}_{\text{uncertainty quantification}} \tag{4}$$

(4) is easier to implement than (2), and is in fact the limit of (2) in the large sample case, because as $n \to \infty$, we have $t_{\alpha/2}(r) \to z_{\alpha/2}$, $g \to 0$, and $\text{Var}(\bar{X}) \sim O(1/n)$. Although Fieller's confidence interval can be more accurate for small samples [30], this benefit appears rather limited for big data. Moreover, the Delta method can also be easily extended for a better approximation by using Edgeworth expansion [5, 29]. To be more specific, (3) suggests that in distribution $\sqrt{n}\left(\bar{Y}/\bar{X} - \mu_y/\mu_x\right)/\sigma_w \to \nu$, whose cumulative distribution function

$$F(t) = \Phi(t) - 6n^{-1/2}\kappa_w(t^2 - 1)\phi(t)$$

contains a correction term in addition to $\Phi(t)$, the cumulative distribution function of the standard normal, and κ_w, the skewness of the W_i's. By simply replacing the terms "$\pm z_{\alpha/2}$" in (4) with $\nu_{\alpha/2}$ and $\nu_{1-\alpha/2}$, the $(\alpha/2)$th and $(1-\alpha/2)$th quantiles of ν, respectively, we obtain the corresponding Edgeworth expansion based confidence interval. Finally, we can add a second-order bias correction term $(\bar{Y}s_x^2/\bar{X} - s_{xy})/(n\bar{X})^2$ to the point estimate in (4); the same correct term can be applied to the Edgeworth expansion based interval.

2.3 Numerical examples

To illustrate the performance of Fieller's interval in (2), the Delta method interval in (4), and the Edgeworth interval with and without bias correction under different scenarios, we let the sample size $n = 20, 50, 200, 2000$. For each fixed n, we assume that the treatment and control groups are independent, and consider three simulation models for i.i.d. experimental units $i = 1, \ldots, n$:

(1) Normal: $X_i \sim N(\mu = 1, \sigma = 0.1)$, $Y_i \sim N(\mu = 1.1, \sigma = 0.1)$;
(2) Poisson: $X_i \sim \text{Pois}(\lambda = 1)$, $Y_i \sim \text{Pois}(\lambda = 1.1)$;
(3) Bernoulli: $X_i \sim \text{Bern}(p = 0.5)$, $Y_i \sim \text{Bern}(p = 0.6)$.

The above models aim to mimic the behaviors of our most common metrics, discrete or continuous. For each case, we repeatedly sample $M = 10,000$ data sets, and for each data set we construct Fieller's and the Delta method intervals, respectively, and then add correction to the Delta method result. We also construct the Edgeworth expansion interval without and with the bias correction.

Table 1: Simulated examples: The first two columns contain simulation models and sample sizes. The next five columns present the coverage rates of various 95% confidence intervals – Fieller's, Delta method based (w/o and w/ bias correction) and Edgeworth expansion based (w/o and w/ bias correction).

Method	n	Fieller	Delta	Delta(BC)	Edgeworth	Edgeworth(BC)
Normal	20	0.9563	0.9421	0.9422	0.9426	0.9426
Normal	50	0.9529	0.9477	0.9477	0.9478	0.9477
Normal	200	0.9505	0.9490	0.9491	0.9490	0.9490
Normal	2000	0.9504	0.9503	0.9503	0.9503	0.9503
Poisson	20	0.9400	0.9322	0.9370	0.9341	0.9396
Poisson	50	0.9481	0.9448	0.9464	0.9464	0.9478
Poisson	200	0.9500	0.9491	0.9493	0.9496	0.9498
Poisson	2000	0.9494	0.9494	0.9495	0.9494	0.9495
Bernoulli	20	0.9539	0.9403	0.9490	0.9476	0.9521
Bernoulli	50	0.9547	0.9507	0.9484	0.9513	0.9539
Bernoulli	200	0.9525	0.9513	0.9509	0.9517	0.9513
Bernoulli	2000	0.9502	0.9500	0.9499	0.9501	0.9500

We report the corresponding coverage rates in Table 1, from which we can draw several conclusions. For big data ($n \geq 200$), all methods achieve nominal (i.e., $\approx$ 95%) coverage rates for all simulation models. For small data ($n \leq 50$), although Fieller's interval seems more accurate for some simulation models (e.g., normal), other methods perform comparably, especially after the bias correction. For simplicity in implementation and transparency in applications, we recommend Algorithm 1, which uses the Delta method based interval (4) with the bias correction.

3 DECODING CLUSTER RANDOMIZATION

3.1 The variance estimation problem

Two key concepts in a typical A/B test are the *randomization unit* – the granularity level where sampling or randomization is performed, and the *analysis unit* – the aggregation level of metric

Algorithm 1 Confidence interval for ratio: Delta method + bias correction

function DELTACI($X = X_1, \ldots, X_n; Y_1, \ldots, Y_n; \alpha = 0.05$)
  $\bar{X} = \text{mean}(X)$; $\bar{Y} = \text{mean}(Y)$;
  $s_x^2 = \text{var}(X)$; $s_y^2 = \text{var}(Y)$; $s_{xy} = \text{cov}(X, Y)$;
  $\text{bc} = \bar{Y}/\bar{X}^3 \times s_x^2/n - 1/\bar{X}^2 \times s_{xy}/n$ ▷ bias correction term
  $\text{pest} = \bar{Y}/\bar{X} - 1 + \text{bc}$ ▷ point estimate
  $\text{vest} = s_y^2/\bar{X}^2 - 2 \times \bar{Y}/\bar{X}^3 * s_{xt} + \bar{Y}^2/\bar{X}^4 * s_x^2$
  return: $\text{pest} \pm z_{1-\alpha/2} \times \sqrt{\text{vest}/n}$ ▷ $100(1-\alpha)$ confidence interval

computation. Analysis is straightforward when the randomization and analysis units agree [14], e.g., when randomizing by user while also computing the average revenue per user. However, often the randomization unit is a *cluster* of analysis units (it cannot be more granular than the analysis unit, otherwise the analysis unit would contain observations under both treatment and control, nullifying the purpose of differentiating the two groups). Such cases, sometimes referred to as cluster randomized experiments in the econometrics and statistics literature [1, 33], are quite common in practice, e.g., enterprise policy prohibiting users within the same organization from different experiences, or the need to reduce bias in the presence of network interference [2, 20, 27]. Perhaps more ubiquitously, for the same experiment we usually have metrics with different analysis units. For example, to meet different business needs, most user-randomized experiments run by ExP contain both user-level and page-level metrics.

We consider two average metrics[5] of the treatment and control groups, assumed to be independent. Without loss of generality, we only focus on the treatment group with K clusters. For $i = 1, \ldots, K$, the ith cluster contains N_i analysis unit level observations Y_{ij} ($j = 1, \ldots, N_i$). Then the corresponding average metric is $\bar{Y} = \sum_{i,j} Y_{ij} \big/ \sum_i N_i$. We assume that within each cluster the observations Y_{ij}'s are i.i.d. with mean μ_i and variance σ_i^2, and across clusters (μ_i, σ_i, N_i) are also i.i.d.

3.2 Existing and Delta method based solutions

$\bar{Y}$ is not an average of i.i.d. random variables, and the crux of our analysis is to estimate its variance. Under strict assumptions, closed-form solutions for this problem exist [19, 33]. For example, when $N_i = m$ and $\sigma_i^2 = \sigma^2$ for all i,

$$\text{Var}(\bar{Y}) = \frac{\sigma^2 + \tau^2}{Km}\{1 + (m-1)\rho\}, \tag{5}$$

where $\tau^2 = \text{Var}(\mu_i)$ is the *between-cluster* variance and $\rho = \tau^2/(\sigma^2 + \tau^2)$ is the *coefficient of intra-cluster correlation*, which quantifies the contribution of between-cluster variance to the total variance. To facilitate understanding of the variance formula (5), two extreme cases are worth mentioning:

(1) If $\sigma = 0$, then for each $i = 1, \ldots, K$ and all $j = 1, \ldots, N_i$, $Y_{ij} = \mu_i$. In this case, $\rho = 1$ and $\text{Var}(\bar{Y}) = \tau^2/K$;
(2) If $\tau = 0$, then $\mu_i = \mu$ for all $i = 1, \ldots, K$, and therefore the observations Y_{ij}'s are in fact i.i.d. In this case, $\rho = 0$ and (5) reduces to $\text{Var}(\bar{Y}) = \sigma^2/(Km)$.

[5]Pedantically, they are two measurements of the same metric. We often use metric to refer to both the metric itself (e.g. revenue-per-user) and measurements of the metric (e.g. revenue-per-user of the treatment group) and this distinction would be clear in the context.

However, unfortunately, although theoretically sound, (5) has only limited practical value because the assumptions it makes are unrealistic. In reality, the cluster sizes N_i and distributions of Y_{ij} for each cluster i are different, which means that μ_i and σ_i^2 are different.

Another common approach is the mixed effect model, also known as multi-level/hierarchical regression [26], where Y_{ij} depends on μ_i and σ_i^2, while the parameters themselves follow a "higher order" distribution. Under this setting, we can infer the treatment effect as the "fixed" effect for the treatment indicator term[6]. *Stan* [9] offers a Markov Chain Monte Carlo (MCMC) implementation aiming to infer the posterior distribution of those parameters, but this needs significant computational effort for big data. Moreover, the estimated ATE, i.e., the coefficient for the treatment assignment indicator, is for the randomization unit (i.e., cluster) but not the analysis unit level, because it treats all clusters with equal weights and can be viewed as the effect on the double average $\sum_i(\sum_j Y_{ij}/N_j)/K$, which is usually different than the population average $\bar{Y}$ [15]. This distinction doesn't make a systematic difference when effects across clusters are homogeneous. However, in practice the treatment effects are often heterogeneous, and using mixed effect model estimates without further adjustment steps could lead to severe biases.

On the contrary, the Delta method solves the problem directly from the metric definition. Re-write $\bar{Y}$ into $\sum_i(\sum_j Y_{ij})/\sum_i N_i$. Let $S_i = \sum_j Y_{ij}$, and divide both the numerator and denominator by K,

$$\bar{Y} = \frac{\sum_i S_i/K}{\sum_i N_i/K} = \bar{S}/\bar{N}.$$

Albeit not an average of i.i.d. random variables, $\bar{Y}$ is a ratio of two averages of i.i.d. randomization unit level quantities [14]. Therefore, by following (3) in Section 2.2,

$$\text{Var}(\bar{Y}) \approx \frac{1}{K\mu_N^2}\left(\sigma_S^2 - 2\frac{\mu_S}{\mu_N}\sigma_{SN} + \frac{\mu_S^2}{\mu_N^2}\sigma_N^2\right). \tag{6}$$

Therefore, the variance of $\bar{Y}$ depends on the variance of a centered version of S_i (by a multiple of N_i, not the constant $\mathbb{E}(S_i)$ as typically done). Intuitively, both the variance of the cluster size N_i, and of the within-cluster sum of observations $S_i = \sum_j Y_{ij}$, contribute to (6). In particular, $\sigma_N^2 = \text{Var}N_i$ is an important contributor of the variance of $\bar{Y}$, leading to a practical recommendation for the experimenters – it is desirable to make the cluster sizes homogeneous; otherwise it can be difficult to detect small treatment effects due to low statistical power.

3.3 Numerical examples

Because the coverage property of (6) has been extensively covered in Section 2.3, we only focus on comparing it with the mixed effect model here. We first describe the data generating process, which consists of a two-level hierarchy as described in the previous section. First, at the randomization unit level, let the total number of clusters $K = 1000$. To mimic cluster heterogeneity, we divide clusters into three categories: small, medium and large. We generate the numbers of clusters in the three categories by the following multinomial

[6]Y ~ Treatment + (1|User) in *lme4* notation. A detailed discussion is beyond the scope of this paper; see Bates et al. [3], Gelman and Hill [26].

distribution:

$$(M_1, M_2, M_3)' \sim \text{Multi-nomial}\{n = K; p = (1/3, 1/2, 1/6)\}.$$

For the ith cluster, depending on which category it belongs to, we generate N_i, μ_i and σ_i in the following way[7]:

- Small: $N_i \sim$ Poisson(2), $\mu_i \sim N(\mu = 0.3, \sigma_i = 0.05)$;
- Medium: $N_i \sim$ Poisson(5), $\mu_i \sim N(\mu = 0.5, \sigma_i = 0.1)$;
- High: $N_i \sim$ Poisson(30), $\mu_i \sim N(\mu = 0.8, \sigma_i = 0.05)$;

Second, for each fixed i, let $Y_{ij} \sim$ Bernoulli($p = \mu_i$) for all $j = 1, \ldots, N_i$. This choice is motivated by binary metrics such as page-click-rate, and because of it we can determine the ground truth $\mathbb{E}(\bar{Y}) = 0.667$ by computing the weighted average of μ_i weighted by the cluster sizes and the mixture of small, medium and large clusters.

Our goal is to infer $\mathbb{E}(\bar{Y})$ and we compare the following three methods:

(1) A naive estimator $\bar{Y}$, pretending all observations Y_{ij} are i.i.d;
(2) Fitting a mixed effect model with a cluster random effect μ_i;
(3) Using the metric $\bar{Y}$ as in the first method, but using the Delta method for variance estimation.

Based on the aforementioned data generating mechanism, we repeatedly and independently generate 1000 data sets. For each data set, we compute the point and variance estimates of $\mathbb{E}(\bar{Y})$ using the naive, mixed effect, and delta methods. We then compute empirical variances for the three estimators, and compare them to the average of estimated variances. We report the results in Table 2.

Table 2: Simulated examples: The first three columns contain the chosen method, the true value of $\mathbb{E}(\bar{Y})$ and the true standard deviation of the corresponding methods. The last two columns contain the point estimates and average estimated standard errors.

Method	Ground Truth	SD(True)	Estimate	Avg. SE(Model)
Naive	0.667	0.00895	0.667	0.00522
Mixed effect	0.667	0.00977	0.547	0.00956
Delta method	0.667	0.00895	0.667	0.00908

Not surprisingly, the naive method under-estimates the true variance – we had treated the correlated observations as independent. Both the Delta method and the mixed effect model produced satisfactory variance estimates. However, although both the naive and the Delta method correctly estimated $\mathbb{E}(Y)$, the mixed effect estimator is severely biased. This shouldn't be a big surprise if we look deeper into the model $Y_{ij} = \alpha + \beta_i + \epsilon_{ij}$ and $\mathbb{E}(\epsilon_{ij}) = 0$, where the random effects β_i are centered so $\mathbb{E}(\beta_i) = 0$. The sum of the intercept terms α and β_i stands for the per-cluster mean μ_i, and α represents the average of per-cluster mean, where we compute the mean within each cluster first, and then average over clusters. This is different from the metrics defined as simple average of Y_{ij} in the way that in the former all clusters are equally weighted and in the latter case bigger clusters have more weight. The two definitions will be the same if and only if either there is no heterogeneity, i.e. per-cluster means μ_i are all the same, or all clusters have the same size. We can still use the mixed effect model to get a unbiased estimate. This requires us to first estimate every β_i (thus μ_i), and then compute $(\alpha + \beta_i)N_i / \sum_i N_i$ by applying the correct weight N_i. The mixed effect model with the above formula gave a new estimate 0.662, much closer to the ground truth. Unfortunately, it is still hard to get the variance of this new estimator.

In this study we didn't consider the treatment effect. In ATE estimation, the mixed effect model will similarly result in a biased estimate for the ATE for the same reason, as long as per-cluster treatment effects vary and cluster sizes are different. The fact that the mixed effect model provides a double average type estimate and the Delta method estimates the "population" mean is analogous to the comparison of the mixed effect model with GEE (generalized estimating equations) [41]. In fact, in the Gaussian case, the Delta method can be seen as the ultimate simplification of GEE's sandwich variance estimator after summarizing data points into sufficient statistics. But the derivation of GEE is much more involved than the central limit theorem, while we can explain the Delta method in a few lines and it is not only more *memorable* but also provides more insights in (6).

4 EFFICIENT VARIANCE ESTIMATION FOR QUANTILE METRICS

4.1 Sample quantiles and their asymptotics

Although the vast majority of metrics are averages of user telemetry data, quantile metrics form another category that is widely used to focus on the tails of distributions. In particular, this is often the case for performance measurements, where we not only care about an average user's experience, but even more so about those that suffer from the slowest responses. Within the web performance community, quantiles (of, for example, page loading time) at 75%, 95% or 99% often take the spotlight. In addition, the 50% quantile (median) is sometimes used to replace the mean, because it is more robust to outlier observations (e.g., errors in instrumentation). This section focuses on estimating the variances of quantile estimates.

Suppose we have n i.i.d. observations $X_1, \ldots, X_n$, generated by a cumulative distribution function $F(x) = P(X \le x)$ and a density function $f(x)$[8]. The theoretical pth quantile for the distribution F is defined as $F^{-1}(p)$. Let $X_{(1)}, \ldots, X_{(n)}$ be the ascending *ordering* of the original observations. The sample quantile at p is $X_{(np)}$ if np is an integer. Otherwise, let $\lfloor np \rfloor$ be the floor of np, then the sample quantile can be defined as any number between $X_{(\lfloor np \rfloor)}$ and $X_{(\lfloor np \rfloor + 1)}$ or a linear interpolation of the two[9]. For simplicity here we use $X_{(\lfloor np \rfloor)}$, which will not affect any asymptotic results. It is a well-known fact that, if $X_1, \ldots, X_n$ are i.i.d. observations, following the central limit theorem and a rather straightforward application of the Delta method, the sample quantile is approximately normal [10, 45]:

$$\sqrt{n}\left\{X_{\lfloor np \rfloor} - F^{-1}(p)\right\} \to N\left[0, \frac{\sigma^2}{f\left\{F^{-1}(p)\right\}^2}\right], \tag{7}$$

where $\sigma^2 = p(1-p)$. However, unfortunately, in practice we rarely have i.i.d. observations. A common scenario is in search engine performance telemetry, where we receive an observation (e.g., page

[7] The positive correlation between μ_i and N_i is not important, and reader can try out code with different configuration.

[8] We do not consider cases when X has discrete mass, and F will have jumps. In this case the quantile can take many values and is not well defined. In practice this case can be seen as continuous case with some discrete correction.

[9] When p is 0.5, the 50% quantile, or median, is often defined as the average of the middle two numbers if we have even number of observations.

loading time) for each server request or page-view, while randomization is done at a higher level such as device or user. This is the same situation we have seen in Section 3, where X_i are clustered. To simplify future notations, we let $Y_i = I\{X_i \le F^{-1}(p)\}$, where I is the indicator function. Then (6) can be used to compute $\mathrm{Var}(\bar{Y})$, and (7) holds in the clustered case with $\sigma^2 = n\mathrm{Var}(\bar{Y})$. This generalizes the i.i.d. case where $n\mathrm{Var}(\bar{Y}) = p(1-p)$. Note that the Delta method is instrumental in proving (7) itself, but a rigorous proof involves a rather technical last step that is beyond our scope. A formal proof can be found in [45].

4.2 A Delta method solution to a practical issue

Although theoretically sound, the difficulty of applying (7) in practice lies in the denominator $f\{F^{-1}(p)\}$, whose computation requires the *unknown* density function f at the *unknown* quantile $F^{-1}(p)$. A common approach is to estimate f from the observed X_i using non-parametric methods such as kernel density estimation [48]. However, any non-parametric density estimation method is trading off between bias and variance. To reduce variance, more aggressive smoothing and hence larger bias need to be introduced to the procedure. This issue is less critical for quantiles at the body of the distribution, e.g. median, where density is high and more data exists around the quantile to make the variance smaller. As we move to the tail, e.g. 90%, 95% or 99%, however, the noise of the density estimation gets bigger, so we have to introduce more smoothing and more bias. Because the density shows up in the denominator and density in the tail often decays to 0, a small bias in estimated density can lead to a big bias for the estimated variance (Brown and Wolfe [7] raised similar criticisms with their simulation study). A second approach is to bootstrap, re-sampling the whole dataset many times and computing quantiles repeatedly. Unlike an average, computing quantiles requires sorting, and sorting in distributed systems (data is distributed in the network) requires data shuffling between nodes, which incurs costly network I/O. Thus, bootstrap works well for small scale data but tends to be too expensive in large scale in its original form (efficient bootstrap on massive data is a research area of its own [34]).

An alternative method without the necessity for density estimation is more desirable, especially from a more practical perspective. One such method is called outer confidence interval (outer CI) [40, 42], which produces a closed-form formula for quantile confidence intervals using combinatorial arguments. Recall that $Y_i = I\{X_i \le F^{-1}(p)\}$ and $\sum Y_i$ is the count of observations no greater than the quantile. In the aforementioned i.i.d. case, $\sum Y_i$ follows a binomial distribution. Consequently, when n is large $\sqrt{n}(\bar{Y} - p) \approx N(0, \sigma^2)$ where $\sigma^2 = p(1-p)$. If the quantile value $F^{-1}(p) \in [X_{(r)}, X_{(r+1)})$, then $\bar{Y} = r/n$. The above equation can be inverted into a $100(1-\alpha)\%$ confidence interval for $r/n : p \pm z_{\alpha/2}\sigma/\sqrt{n}$. This means with 95% probability the true percentile is between the lower rank $L = n(p - z_{\alpha/2}\sigma/\sqrt{n})$ and upper rank $U = n(p + z_{\alpha/2}\sigma/\sqrt{n}) + 1$!

The traditional outer CI depends on X_i being i.i.d. But when there are clusters, σ^2/n instead of being $p(1-p)/n$ simply takes a different formula (6) by the Delta method and the result above still holds. Hence the confidence interval for a quantile can be computed in the following steps:

(1) fetch the quantile $X_{(\lfloor np \rfloor)}$
(2) compute $Y_i = I\{X_i \le X_{(\lfloor np \rfloor)}\}$
(3) compute μ_S, μ_N, σ_S^2, σ_{SN}, σ_N^2
(4) compute σ by setting σ^2/n equal to the result of equation (6)
(5) compute $L, U = n(p \pm z_{\alpha/2}\sigma)$
(6) fetch the two ranks $X_{(L)}$ and $X_{(U)}$

We call this outer CI with pre-adjustment. This method reduces the complexity of computing a quantile and its confidence interval into a Delta method step and subsequently fetching three "ntiles". However, in this algorithm the ranks depends on σ, whose computation depends on the quantile estimates (more specifically the Y_i requires a pass through the data after quantile is estimated). This means that this algorithm requires a first 'ntile' retrieval, and then a pass through the data for σ computation, and then another two 'ntile' retrievals. Turns out, computing all three 'ntiles' in one stage is much more efficient than splitting into two stages. This is because retrieving 'ntiles' can be optimized in the following way: if we only need to fetch tail ranks, it is pointless to sort data that are not at the tail; we can use sketching algorithm to narrow down the possible range where our ranks reside and only sort in that range, making it even more efficient to retrieve multiple 'ntiles' at once. Along this line of thoughts, to make the algorithm more efficient, we noticed that (7) also implies that the change from X_i being i.i.d. to clustered only requires an adjustment to the numerator σ, which is a simple re-scaling step, and the correction factor does not depend on the unknown density function f in the denominator. If the outer CI were to provide a good confidence interval in i.i.d. case, a re-scaled outer CI with the same correction term should also work for the clustered case, at least when n is large. This leads to the outer CI with post-adjustment algorithm:

(1) compute $L, U = n(p \pm z_{\alpha/2}\sqrt{p(1-p)/n})$
(2) fetch $X_{(\lfloor np \rfloor)}$, $X_{(L)}$ and $X_{(U)}$
(3) compute $Y_i = I\{X_i \le X_{(\lfloor np \rfloor)}\}$
(4) compute μ_S, μ_N, σ_S^2, σ_{SN}, σ_N^2
(5) compute σ by setting σ^2/n equal to the result of equation (6)
(6) compute the correction factor $\sigma/\sqrt{p(1-p)}$ and apply it to $X_{(L)}$ and $X_{(U)}$

We implemented this latter method in ExP using Apache Spark [51] and SCOPE [11].

4.3 Numerical examples

To test the validity and performance of the adjusted outer CI method, we compare its coverage to a standard non-parametric bootstrap ($N_B = 1000$ replicates). The simulation setup consists of $N_u = 100, \ldots, 10000$ users (clusters) with $N_i^u = 1, \ldots, 10$ observations each (N_i^u are uniformly distributed). Each observation is the sum of two i.i.d. random variables $X_i^u = X_i + X_u$, where X_u is constant for each user. We consider two cases, one symmetric and one heavy-tailed distribution:

- Normal: $X_i, X_u \overset{\text{iid}}{\sim} \mathcal{N}(\mu = 0, \sigma = 1)$;
- Log-Normal: $X_i, X_u \overset{\text{iid}}{\sim}$ Log-normal$(\mu = 0, \sigma = 1)$.

First, we find the "true" 95th percentile value of these distribution by computing its value for a very large sample ($N_u = 10^7$). Second, we compute the confidence intervals for $M = 10000$ simulation runs

using bootstrap and outer CI with pre- and post-adjustment and compare their coverage estimates (≈0.002 standard error), shown in Table 3. We found that when the sample contains 1000 or more clusters, all methods provide good coverage. Pre- and post-adjustment outer CI results are both very close to the much more computationally expensive bootstrap (in our un-optimized simulations, the outer CI method was ≈20 times faster than bootstrap). When the sample size was smaller than 1000 clusters, bootstrap was noticeably inferior to outer CI. For all sample sizes, pre-adjustment provided slightly larger coverage than post-adjustment, and this difference increased for smaller samples. In addition, because adjustment tends to result in increased confidence intervals, unadjusted ranks are more likely to have the same value as the quantile value, and thus post-adjustment is more likely to underestimate the variance in that case. In conclusion, post-adjustment outer CI works very well for large sample sizes $N_u \geq 1000$ and reasonably well for smaller samples, but has slightly inferior coverage compared to pre-adjustment. For big data, outer CI with post-adjustment is recommended due its efficiency, while for median to small sample sizes, outer CI with pre-adjustment is preferred if accuracy is paramount.

Table 3: Simulated examples: The first two columns contain the simulated models and sample sizes. The last three columns contain the coverage rates for the Bootstrap, pre-adjusted and post-adjusted outer confidence intervals.

Distribution	N_u	Bootstrap	Outer CI (pre-adj.)	Outer CI (post-adj.)
Normal	100	0.9039	0.9465	0.9369
Normal	1000	0.9500	0.9549	0.9506
Normal	10000	0.9500	0.9500	0.9482
Log-normal	100	0.8551	0.9198	0.9049
Log-normal	1000	0.9403	0.9474	0.9421
Log-normal	10000	0.9458	0.9482	0.9479

5 MISSING DATA AND WITHIN-SUBJECT ANALYSES

5.1 Background

Consider the case that we are tracking a metric over time. For example, businesses need to track key performance indicators like user engagement and revenue daily or weekly for a given audience. To simplify, let us say we are tracking weekly visits-per-user. Visits-per-user is defined as a simple average $\bar{X}_t, t = 1, \ldots$ for each week. If there is a drop between $\bar{X}_t$ and $\bar{X}_{t-1}$, how do we set up an alert so that we can avoid triggering it due to random noise and control the false alarm rate? Looking at the variance, we see

$$\mathrm{Var}(\bar{X}_t - \bar{X}_{t-1}) = \mathrm{Var}(\bar{X}_t) + \mathrm{Var}(\bar{X}_{t-1}) - 2\mathrm{Cov}(\bar{X}_t, \bar{X}_{t-1})$$

and we need to have a good estimate of the covariance term because we know it is non-zero.

Normally, estimating the covariance of two sample averages is trivial and the procedure is very similar to the estimation of the sample variance. But there is something special about this case — missing data. Not everyone uses the product every week. For any online and offline metric, we are only able to define the metric using observed data. If for every user we observe its value in week $t-1$ and t, the covariance can be estimated using the sample covariance formula. But if there are many users who appear in week $t-1$ but not in t, it is unclear how to proceed. The naive approach is to estimate the covariance using complete observations, i.e. users who appear in both weeks. However, this only works if the data is *missing completely at random*. In reality, active users will show up more often and are more likely to appear in both weeks, meaning the missing data are obviously not random.

In this section we show how the Delta method can be very useful for estimating $\mathrm{Cov}(\bar{X}_t, \bar{X}_{t'})$ for any two time points t and t'. We then use this to show how we can analyze within-subject studies, also known as pre-post, repeated measurement, or longitudinal studies. Our method starts with metrics directly, highly contrasting with traditional methods such as mixed effect models which build models from individual user's data. Because we study metrics directly, our model is *small* and easy to solve with its complexity constant to the scale of data.[10]

5.2 Methodology

How do we estimate covariance with a completely unknown missing data mechanism? There are many existing works on handling missing data. One approach is to model the propensity of a datapoint being missing using other observed predictors [13]. This requires additional covariates/predictors to be observed, plus a rather strong assumption that conditioned on these predictors, data are missing completely at random. We present a novel idea using the Delta method after *data augmentation* without the need of modeling the missing data mechanism. Specifically, we use an additional indicator for the presence/absence status of a user in each period t. For user i in period t, let $I_{it} = 1$ if user i appears in period t, and 0 otherwise. For each user i in period t, instead of one scalar metric value (X_{it}), we augment it to a vector (I_{it}, X_{it}). When $I_{it} = 0$, i.e. user is missing, we define $X_{it} = 0$. Under this simple augmentation, the metric value $\bar{X}_t$, taking the average over those non-missing measurements in period t, is the same as $\sum_i X_{it} / \sum_i I_{it}$! In this connection,

$$\mathrm{Cov}(\bar{X}_t, \bar{X}_{t'}) = \mathrm{Cov}\left(\frac{\sum_i X_{it}}{\sum_i I_{it}}, \frac{\sum_i X_{it'}}{\sum_i I_{it'}}\right) = \mathrm{Cov}\left(\frac{\bar{X}_t}{\bar{I}_t}, \frac{\bar{X}_{t'}}{\bar{I}_{t'}}\right)$$

where the last equality is by dividing both numerator and denominator by the same total number of users who have appeared in any of the two periods.[11] Thanks to the central limit theorem, the vector $(\bar{I}_t, \bar{X}_t, \bar{I}_{t'}, \bar{X}_{t'})$ is also asymptotically (multivariate) normal with covariance matrix Σ, which can be estimated using sample variance and covariance because there is *no missing data* after augmentation. In Section 2 we already applied the Delta method to compute the variance of a ratio of metrics by taking the first order Taylor expansion. Here, we can expand the two ratios to their first order linear form $(\bar{X}_t - \mu_{X_t})/\mu_{I_t} - \mu_{X_t}(\bar{I}_t - \mu_{I_t})/\mu_{I_t}^2$, where μ_X and μ_I are the means of X and I and the expansion for t' is similar. $\mathrm{Cov}(\bar{X}_t, \bar{X}_{t'})$ can then be computed using Σ.

[10] Part of the work in this section has previously appeared in a technical report [28]. Since authoring the technical report, the authors developed a better understanding of the differences between mixed effect models and the proposed method, and we present our new findings here. The technical report has more details about other types of repeated measurement models.

[11] Actually, if there are more than two period, we can either use only users appeared in any of the two periods, or users appeared in any of all the periods. It is mathematically the same thing if we added more users and then treat them as not appeared in I and X, i.e. $\bar{X}_t$ remains the same.

We can apply this to the general case of a within-subject study. Without loss of any generality and with the benefit of a concrete solution, we only discuss a cross-over design here. Other designs including more periods are the same in principle. In a cross-over design, we have two groups I and II. For group I, we assign them to the treatment for the first period and control for the second. For group II, the assignment order is reversed. We often pick the two groups using random selection, so each period is an A/B test by itself.

Let $\bar{X}_{it}, i = 1, 2, t = 1, 2$ be the metric for group i in period t. Let $\mathbf{X} = (\bar{X}_{11}, \bar{X}_{12}, \bar{X}_{21}, \bar{X}_{22})$. We know $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma_{\mathbf{X}})$ for a mean vector $\boldsymbol{\mu}$ and covariance matrix $\Sigma_{\mathbf{X}}$. $\Sigma_{\mathbf{X}}$ has the form $\text{diag}(\Sigma, \Sigma)$ since the two groups are independent with the same distribution. With the help of the Delta method, we can estimate Σ from the data and treat it as a constant covariance matrix (hence $\Sigma_{\mathbf{X}}$). Our model concerns the mean vector $\boldsymbol{\mu}(\boldsymbol{\theta})$ using other parameters which represent our interest. In a cross-over design, $\boldsymbol{\theta} = (\theta_1, \theta_2, \Delta)$ where the first two are baseline means for the two periods and our main interest is the treatment effect Δ. (We assume there is no treatment effect for group I carried over to the 2nd period. To study carry over effect, a more complicated design needs to be employed.) In this parameterization,

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = (\theta_1 + \Delta, \theta_2, \theta_1, \theta_2 + \Delta)\ . \tag{8}$$

The maximum likelihood estimator and Fisher information theory [45] paved a general way for us to estimate $\boldsymbol{\theta}$ as well as the variance of the estimators for various mean vector models. For example, if we want to model the relative change directly, we just need to change the addition of Δ into multiplication of $(1 + \Delta)$. Notice the model we are fitting here is very small, almost a toy example from a text book. All the heavy lifting is in computing $\Sigma_{\mathbf{X}}$ which is dealt with by the Delta method where all computations are trivially distributed. When $\boldsymbol{\mu}(\boldsymbol{\theta}) = M\boldsymbol{\theta}$ is linear as in the additive cross-over model above, $\hat{\boldsymbol{\theta}} = (M^T\Sigma_{\mathbf{X}}^{-1}M)^{-1}M^T\Sigma_{\mathbf{X}}^{-1}\mathbf{X}$ and $\text{Var}(\widehat{\boldsymbol{\theta}}) = I^{-1}$ where the Fisher Information $I = M^T\Sigma_{\mathbf{X}}^{-1}M$.

5.3 Numerical examples

We simulate two groups with 1000 users each. The first group receives treatment in period 1, then control in period 2, while the second group receives the inverse. For each user, we impose an independent user effect u_i that follows a normal distribution with mean 10 and standard deviation 3, and independent noises ϵ_{it} with mean 0 and standard deviation 2. Each user's base observations (before treatment effect) for the two periods are $(u_i + \epsilon_{i1}, u_i + \epsilon_{i2})$. We then model the missing data and treatment effect such that they are correlated. We define a user's engagement level l_i by its user effect u_i through $P(U < u_i)$, i.e. the probability that a user's u is bigger than another random user. We model the treatment effect Δ_i as an additive normal with mean 10 and standard deviation 0.3, multiplied by l_i. For each user and each of the two periods, there is a 1 - max(0.1, l_i) probability of this user being missing. We can interpret the missing data as a user not showing up, or as a failure to observe. In this model, without missing data, every user has two observations and the average treatment effect should be $E(\Delta_i) = 5$ because $E(l_i) = 0.5$. Since we have missing data and it is more likely for lower engagement levels to be missing, we expect the average treatment effect for the population of all observed users to be between 5 to 10. In the current model we didn't add a time period effect such that the second period could have a different mean θ_2 from the first period's mean θ_1, but in analysis we always assume that this effect could exist.

We simulate the following 1000 times: each time we run both the mixed effect model $X_{it} \sim$ IsTreatment + Time + (1|User) as well as the additive cross-over model (8) and record their estimates of the ATE and the corresponding estimated variance. We then estimate the true estimator variance using the sample variance of those estimates among 1000 trials and compare that to the mean of the estimated variance to evaluate the quality of variance estimations. We also compute the average of ATE estimates and compare to the true ATE to assess the bias.

Table 4: Simulated examples: The first three columns contain the method, true ATE and standard deviations of the corresponding methods. The last two columns contain the point estimates and average estimated standard errors.

	Ground Truth	SD(True)	Estimate	Avg. SE(Model)
mixed effect	6.592	0.1295	7.129	0.1261
Delta method	6.592	0.1573	6.593	0.1568

Table 4 summarizes the results. We found both methods provide good variance estimation and the mixed effect model shows smaller variance. However, mixed effect also displays an upward bias in the ATE estimate while the Delta method closely tracks the true effect. To further understand the difference between the two, we separate the data into two groups: users with complete data, and users who only appear in one period (incomplete group). We run the mixed effect model for the the two groups separately. Note that in the second group each user only appears once in the data, so the model is essentially a linear model. Our working hypothesis is the following: because the mixed effect model assumes a fixed treatment effect, the effect for the complete and incomplete groups must be the same. The mixed effect model can take advantage of this assumption and construct an estimator by weighted average of the two estimates from the two groups, with the optimal weight inversely proportional to their variances. Table 5 shows the weighted average estimator is indeed very close to mixed effect model for both estimate and variance. The weighted average is closer to the complete group because the variance there is much smaller than that of the incomplete group since within-subject comparison significantly reduces noise. This explains why the mixed effect model can produce misleading results in within-subject analysis whenever missing data patterns can be correlated with the treatment effect. The Delta method, on the other hand, offers a flexible and robust solution.

Table 5: Simulated examples: Point and variance estimates using the mixed effect vs. weighted average estimators.

	Estimate	Var
mixed effect model	7.1290	0.00161
mixed effect model on complete group	7.3876	0.00180
linear model on incomplete group	5.1174	0.01133
weighted avg estimator	7.0766	0.00155

6 CONCLUDING REMARKS

Measure everything is not only an inspiring slogan, but also a crucial step towards the holy grail of data-driven decision making. In the

big data era, innovative technologies have tremendously advanced user telemetry and feedback collection, and distilling insights and knowledge from them is an imperative task for business success. To do so, we typically apply certain statistical models at the level of individual observations, fit the model using numerical procedures such as solving optimization problems, and draw conclusions and assess uncertainties from the fitted models. However, for big data such an analytical procedure can be very challenging. Based on the key observation that metrics are approximately normal due to the central limit theorem, this paper offers an alternative perspective by advocating modeling metrics, i.e., summaries or aggregations of the original data sets, in a direct manner. By doing so, we can decompose big data into small problems. However, although conceptually sound, in practice these metric level models often involve nonlinear transformations of data or complex data generating mechanisms, posing several challenges for trustworthy and efficient metric analytics.

To address these issues, we promoted the Delta method's central role in making it possible to extend the central limit theorem to new territories. We demonstrated how to apply the Delta method in four real-life applications, and illustrated how this approach naturally leads to trivially parallelizable and highly efficient implementations. Among these applications, ratio metrics, clustered randomizations and quantile metrics are all common and important scenarios for A/B testing, and business analytics in general. Within-subject studies are becoming more popular for superior sensitivities, and missing data with an unknown mechanism is ubiquitous in both the online and offline worlds. Our contribution to technical improvements, novel ideas and new understandings includes bias correction for ratio metric estimation, combination of the Delta method and outer confidence intervals for quantile metric variance estimation, and the idea of data augmentation for general missing data problems in within-subject studies. We also revealed the connection between the Delta method and mixed effect models, and explained their differences. In addition, we pointed out the advantage of the Delta method in the presence of an unknown missing data mechanism. Overall speaking, we hope this paper can serve as a practical guide of applying the Delta method in large-scale metric analyses and A/B tests, so that it is no longer just a technical detail, but a starting point and a central piece of analytical thinking.

Although the Delta method can help tackle big data problems, it does not replace the need for rigorous experimental designs and probabilistic modeling. For example, "optimal" choices for cluster configurations, randomization mechanisms and data transformations are known to increase the sensitivities of metrics [8, 32, 49]. We leave them as future research directions.

ACKNOWLEDGMENTS

We benefited from insightful discussions with Ya Xu and Daniel Ting on quantile metrics and outer confidence intervals. We thank Jong Ho Lee for implementing the efficient quantile metric estimation and confidence interval construction in ExP using Apache Spark, and Yu Guo for previous work on repeated measurements.

REFERENCES

[1] Susan Athey and Guido W Imbens. 2017. The econometrics of randomized experiments. *Handbook of Economic Field Experiments* 1 (2017), 73–140.
[2] Lars Backstrom and Jon Kleinberg. 2011. Network bucket testing. In *Proceedings of the 20th international conference on World wide web*. ACM, 615–624.
[3] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823* (2014).
[4] Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, et al. 2014. lme4: Linear mixed-effects models using Eigen and S4. *R package version* 1, 7 (2014), 1–23.
[5] Dennis D Boos and Jacqueline M Hughes-Oliver. 2000. How large does n have to be for Z and t intervals? *The American Statistician* 54, 2 (2000), 121–128.
[6] Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Springer, 177–186.
[7] Morton B Brown and Robert A Wolfe. 1983. Estimation of the variance of percentile estimates. *Computational Statistics & Data Analysis* 1 (1983), 167–174.
[8] Roman Budylin, Alexey Drutsa, Ilya Katsev, and Valeriya Tsoy. 2018. Consistent Transformation of Ratio Metrics for Efficient Online Controlled Experiments. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 55–63.
[9] Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Michael A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2016. Stan: A probabilistic programming language. *Journal of Statistical Software* 20 (2016), 1–37.
[10] George Casella and Roger L Berger. 2002. *Statistical Inference, Second Edition*. Duxbury Press: Pacific Grove, CA.
[11] Ronnie Chaiken, Bob Jenkins, Per-Åke Larson, Bill Ramsey, Darren Shakib, Simon Weaver, and Jingren Zhou. 2008. SCOPE: Easy and efficient parallel processing of massive data sets. *Proceedings of the VLDB Endowment* 1 (2008), 1265–1276.
[12] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20 (1995), 273–297.
[13] M. Davidian, A.A. Tsiatis, and S. Leon. 2005. Semiparametric Estimation of Treatment Effect in a Pretest-Posttest Study with Missing Data. *Statist. Sci.* 20 (2005), 295–301. Issue 3.
[14] A. Deng, J. Lu, and J. Litz. 2017. Trustworthy analysis of online A/B tests: Pitfalls, challenges and solutions. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 641–649.
[15] A. Deng and X. Shi. 2016. Data-driven metric development for online controlled experiments: Seven lessons learned. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
[16] Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. 2013. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the 6th ACM WSDM Conference*. 123–132.
[17] Pavel Dmitriev, Somit Gupta, Dong Woo Kim, and Garnet Vaz. 2017. A Dirty Dozen: Twelve Common Metric Interpretation Pitfalls in Online Controlled Experiments. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. ACM, New York, NY, USA, 1427–1436. https://doi.org/10.1145/3097983.3098024
[18] Pavel Dmitriev and Xian Wu. 2016. Measuring Metrics. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 429–437.
[19] Allan Donner. 1987. Statistical methodology for paired cluster designs. *American Journal of Epidemiology* 126, 5 (1987), 972–979.
[20] Dean Eckles, Brian Karrer, and Johan Ugander. 2017. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference* 5, 1 (2017).
[21] Jianqing Fan, Fang Han, and Han Liu. 2014. Challenges of big data analysis. *National Science Review* 1 (2014), 293–314.
[22] Edgar C Fieller. 1940. The biological standardization of insulin. *Supplement to the Journal of the Royal Statistical Society* 7, 1 (1940), 1–64.
[23] Edgar C Fieller. 1954. Some problems in interval estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* (1954), 175–185.
[24] Ronald Aylmer Fisher. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 222 (1922), 309–368.
[25] Pedro A Forero, Alfonso Cano, and Georgios B Giannakis. 2010. Consensus-based distributed support vector machines. *Journal of Machine Learning Research* 11, May (2010), 1663–1707.
[26] Andrew Gelman and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.
[27] Huan Gui, Ya Xu, Anmol Bhasin, and Jiawei Han. 2015. Network A/B testing: From sampling to estimation. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 399–409.
[28] Yu Guo and Alex Deng. 2015. Flexible Online Repeated Measures Experiment. *arXiv preprint arXiv:1501.00450* (2015).
[29] Peter Hall. 2013. *The bootstrap and Edgeworth expansion*. Springer Science & Business Media.
[30] Joe Hirschberg and Jenny Lye. 2010. A geometric comparison of the delta and Fieller confidence intervals. *The American Statistician* 64 (2010), 234–241.

[31] Michael I Jordan, Jason D Lee, and Yun Yang. 2018. Communication-efficient distributed statistical inference. *J. Amer. Statist. Assoc.* in press (2018). doi:10.1080/01621459.2018.1429274.

[32] Eugene Kharitonov, Alexey Drutsa, and Pavel Serdyukov. 2017. Learning sensitive combinations of a/b test metrics. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 651–659.

[33] Neil Klar and Allan Donner. 2001. Current and future challenges in the design and analysis of cluster randomization trials. *Statistics in medicine* 20, 24 (2001), 3729–3740.

[34] Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I Jordan. 2014. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76, 4 (2014), 795–816.

[35] Ronny Kohavi, Thomas Crook, Roger Longbotham, Brian Frasca, Randy Henne, Juan Lavista Ferres, and Tamir Melamed. 2009. Online experimentation at Microsoft. In *Proceedings of the Third International Workshop on Data Mining Case Studies, held at the 5th ACM SIGKDD Conference*. 11–23.

[36] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online Controlled Experiments at Large Scale. *Proceedings of the 19th ACM SIGKDD Conference* (2013).

[37] Ron Kohavi, Randal M Henne, and Dan Sommerfield. 2007. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *Proceedings of the 13th ACM SIGKDD Conference*. 959–967.

[38] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery* 18, 1 (2009), 140–181.

[39] R. Kohavi, R. Longbotham, and T. Walker. 2010. Online Experiments: Practical Lessons. *Computer* 43, 9 (Sept 2010), 82–85.

[40] Daniel Krewski. 1976. Distribution-free confidence intervals for quantile intervals. *J. Amer. Statist. Assoc.* 71, 354 (1976), 420–422.

[41] Kung-Yee Liang and Scott L Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73, 1 (1986), 13–22.

[42] John S Meyer. 1987. Outer and inner confidence intervals for finite population quantile intervals. *J. Amer. Statist. Assoc.* 82, 397 (1987), 201–204.

[43] Walter Rudin et al. 1964. *Principles of mathematical analysis*. Vol. 3. McGraw-hill New York.

[44] Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. 2010. Overlapping Experiment Infrastructure: More, Better, Faster Experimentation. *Proceedings of the 16th ACM SIGKDD Conference* (2010).

[45] Aad W Van der Vaart. 2000. *Asymptotic statistics*. Vol. 3. Cambridge university press.

[46] Ulrike Von Luxburg and Volker H Franz. 2009. A geometric approach to confidence sets for ratios: Fieller's theorem, generalizations and bootstrap. *Statistica Sinica* (2009), 1095–1117.

[47] Dongli Wang and Yan Zhou. 2012. Distributed support vector machines: An overview. In *Control and Decision Conference (CCDC), 2012 24th Chinese*. IEEE, 3897–3901.

[48] Larry Wasserman. 2003. *All of Statistics: A Concise Course in Statistical Inference*. Springer.

[49] Huizhi Xie and Juliette Aurisset. 2016. 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*. ACM, 645–654.

[50] Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin. 2015. From infrastructure to culture: A/B testing challenges in large scale social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2227–2236.

[51] Matei Zaharia, Reynold S Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J Franklin, et al. 2016. Apache Spark: A unified engine for big data processing. *Commun. ACM* 59, 11 (2016), 56–65.

[52] Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J Smola. 2010. Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems*. 2595–2603.