

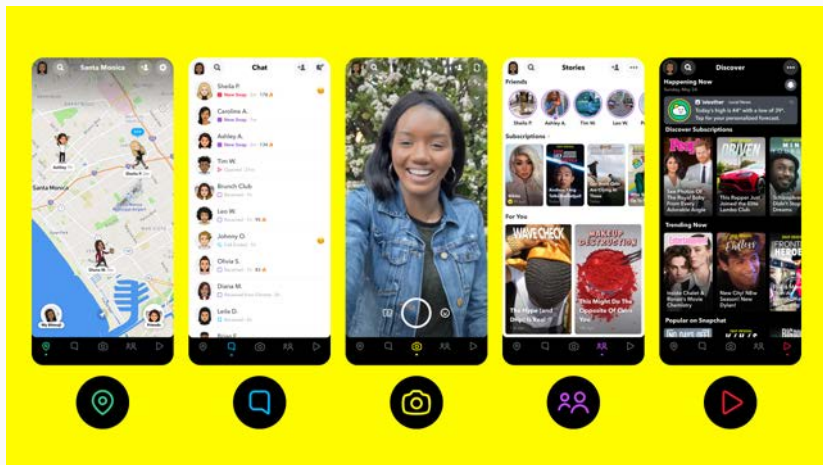
CONQ: CONTinuous Quantile Treatment Effects for Large-Scale Online Controlled Experiments

Weinan Wang, Xi Zhang

Snap Inc.

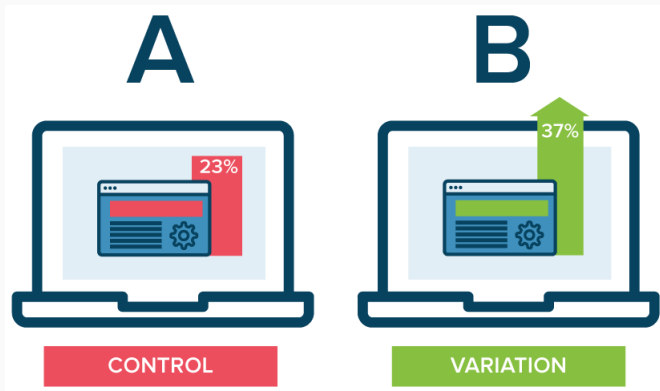
Introduction

Snapchat is a camera company:



Introduction

We have over 2000 metrics measuring our customers' engagement with the app (engagement metrics: active days, session time, #chats, #snaps, #story views, #story posts, ...), together with the app's performance (latency metrics: camera load latency, page load latency, etc.).



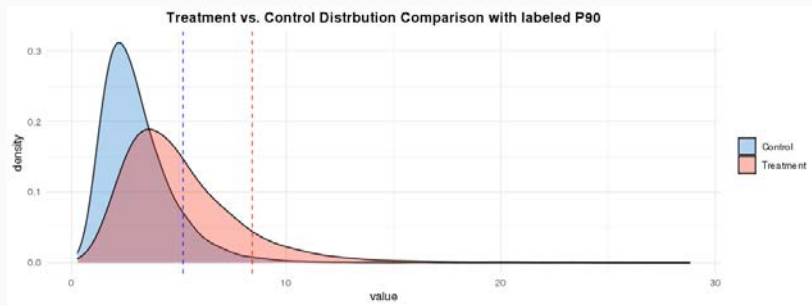
Experimentation

In A/B tests, we measure changes in engagement at mean using two-sample t -test for each metric.

Each user contributes exactly one data point, since randomization units are also users, they can be considered i.i.d.

Background

For performance metrics, the industry standard is to measure event level QTE (quantile treatment effects), most notably at P50 and P90.



- **each user contributes multiple events, where they are no longer independent.** Highly engaged users would contribute more data points, whereas rarely engaged users would contribute less. So event level performance metrics overweight power users.

Challenges

- Given such user level dependency, **delta-method** is the currently wide-adopted approach in calculating QTE at given quantiles, it's cumbersome to generalize to a continuous range. For example, in order to calculate QTE at P50, each user i as P_i event contributing to the quantile and have within user correlation, so:

$$Y_n(x) = \frac{Z_1 + \dots + Z_n}{P_1 + \dots + P_n}$$




Where Z is the number of events greater than P50 for each user. So by delta method,

$$\text{Var}(Y_n(x)) = \frac{1}{n} \left(\frac{\text{Var}(Z)}{\bar{p}^2} + \frac{\text{Var}(P)\bar{Z}^2}{\bar{p}^4} - 2\frac{\bar{Z}}{\bar{p}^3} \text{Cov}(Z, P) \right)$$

- We further need $f(x)$ term in variance calculation at quantiles. However, we have little guidance on density estimation's bandwidth choice, which greatly affects the final result.

Challenges

Furthermore, QTE at the median and the 90-th quantile are sometimes not enough to give experimenters the whole picture, especially when the significance or directions do not agree, or when heterogeneous treatment effect (HTE) is present for different devices with high and low overall performances.

Metrics		Absolute Change	% Change
Latency P50	 	1,540.833 → 1,545.322	0.29% ±0.07% p < 0.001
Latency P90		6,601.484 → 6,592.833	-0.13% ±0.07% p < 0.001

CONQ: Main Contributions

- A scalable and theoretically sound method that can provide QTE with $\Delta\%$ p -values and CIs at arbitrary range of quantile locations simultaneously.
- Circumvents the issue of density estimation and bandwidth choice altogether, improve accuracy of delta-method by log-transformation for percentage changes.
- Validation of the method on real experiments at Snap Inc., which shows consistency with existing P50 and P90 results, and stable performance across various quantile locations.

Methodology

Sample Quantiles

CLT for i.i.d. sample quantiles

For i.i.d. samples X_1, \dots, X_n , the sample p -th quantile $\xi_{n,p}$ is asymptotically normal:

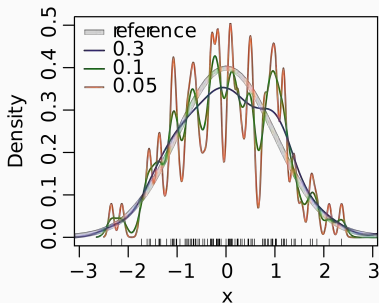
$$\sqrt{n} (\xi_{n,p} - \xi_p) \rightarrow \mathcal{N} \left(0, \frac{p(1-p)}{f(\xi_p)^2} \right)$$

However, for performance metrics, each user can contribute multiple data-points (where itself is another r.v. measuring engagement), and observations are no longer i.i.d.; e.g., a power Snapchat user's latency events would be highly correlated with each other given if the user uses a single iPhone 12.

Sample Quantiles

The industry standard is to use the delta-method to approximate the numerator in the variance term [2, 4] as mentioned before. Yet we still have $f(\xi_p)$ to estimate, with kernel density estimation as the standard technique. However, it's extremely sensitive to the choice of bandwidth h .

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$



Kernel Density Bandwidths

Here is an example on a typical performance metric measuring the latency on starting the Discover section inside the Snapchat app. The study improved P50 latency from 911.991 to 908.173 with a p -value of 0.044 using the recommended bandwidth choice in [4]. However:

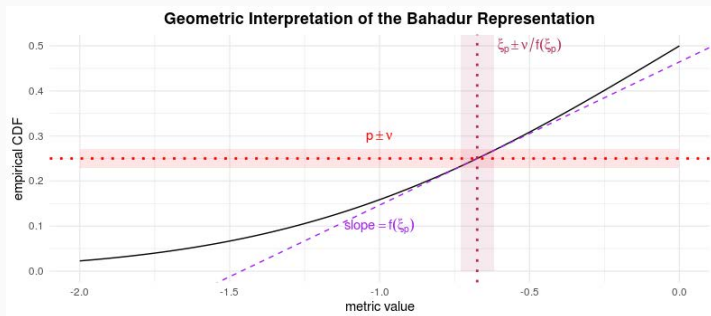
kernel density estimator's bandwidth	$\Delta\%$ p -value
0.002	1.08e-34
0.01	0.0125
0.02	0.044
normal reference rule ¹	0.352

¹[5], chosen as $h_n = 1.06 \min\{s, \text{IQR}/1.34\}/n^{1/5}$ (IQR inter-quartile range, s is sample standard deviation and n is sample size)

Bahadur Representation

Luckily, we do have a relationship between quantiles and the CDF (even for some dependent cases under regularity conditions [7]):

$$\xi_{n,p} = \xi_p + \frac{p - F_n(\xi_p)}{f(\xi_p)} + O_{a.s.}[n^{-3/4}(\log n)^{1/2}(\log \log n)^{1/4}]$$



i.i.d. samples to dependent samples

i.i.d. samples

$$\sqrt{n}(\xi_{n,p} - \xi_p) \rightarrow \mathcal{N}\left(0, \frac{p(1-p)}{f(\xi_p)^2}\right) \Rightarrow p - F_n(\xi_p) \rightarrow \mathcal{N}\left(0, \frac{p(1-p)}{n}\right)$$

dependent samples

$$\xi_{n,p} - \xi_p \rightarrow \mathcal{N}\left(0, \frac{\nu}{f(\xi_p)^2}\right) \Rightarrow p - F_n(\xi_p) \rightarrow \mathcal{N}(0, \nu)$$

Woodruff Confidence Interval

Woodruff [6] further proposed the empirical CI for quantiles by inverting the usual confidence intervals for the distribution function:

$$\text{Let } \hat{\xi}_p^L = \inf \left\{ t : F_n(t) \geq p - z_{\alpha/2} \nu^{1/2} [F_n(\xi_p)] \right\},$$

$$\hat{\xi}_p^U = \inf \left\{ t : F_n(t) \geq p + z_{\alpha/2} \nu^{1/2} [F_n(\xi_p)] \right\}, \text{ then}$$

$$\mathbb{P} \left(\hat{\xi}_p^L \leq \xi_p \leq \hat{\xi}_p^U \right) \approx 1 - \alpha.$$

Note: Woodruff CI need not be symmetric against ξ_p , and especially so at large or small value of p . So we choose $\max\{\hat{\xi}_p^U - \hat{\xi}_p, \hat{\xi}_p - \hat{\xi}_p^L\}$ as conservative estimate of the standard error.

So to get QTE for performance metrics, we

- shift our focus from sample quantiles to empirical distributions, which avoids density estimation
- estimate variance term using bootstrap to circumvent dependency
- appeal to Woodruff CI to get conservative estimates of the standard errors for both control and treatment
-

Balanced BLB (Poorman's Bootstrap) for CDF

In order to estimate ν , which is the key quantity for conducting point-wise statistical inference using the aforementioned paradigm, we propose a balanced poorman's version of the famed **Bag-of-Little-Bootstrap (BLB) [3]** procedure on the CDF.

- So instead of bootstrapping on users which is computationally extensive, we bootstrap on bucketized aggregated summary statistics.
- Events that belong to the same user are in the same bucket to ensure validity.

Step 1: log-transformation

Due to the non-negative nature of latency metrics and quantile's invariance to monotone transformation [1], we have:

$$\log \xi_{f(x)}(p) = \xi_{f(\log x)}(p), \forall 0 < p < 1.$$

We do this to make delta method more appropriate for Taylor expansion, and also reduce the overall unique value of log-latency metric (after rounding).

Step 1: log-transformation

This ensures that we can get back the QTE on the original metric X by simply taking the exponent, and by delta-method, we also have:

variance on $\Delta\%$ between control and treatment at p -th quantile

$$\left(\text{Var} \left(\xi_{f(\log x)}^T(p) \right) + \text{Var} \left(\xi_{f(\log x)}^C(p) \right) \right) \left(\frac{\xi_{f(x)}^T(p)}{\xi_{f(x)}^C(p)} \right)^2$$

where the quantity in the first bracket can be estimated separately using our bootstrapping procedure.

Step 2: bootstrapping on buckets

- round log-transformed metrics into desired precision (2 digits for Snap), \tilde{X} .
- split N users together with their events into s subsets ($s = 100$ for Snap), and for each bucket, record the unique metric values and their corresponding aggregated counts.

Step 2: bootstrapping on buckets

- randomly permute the long vector

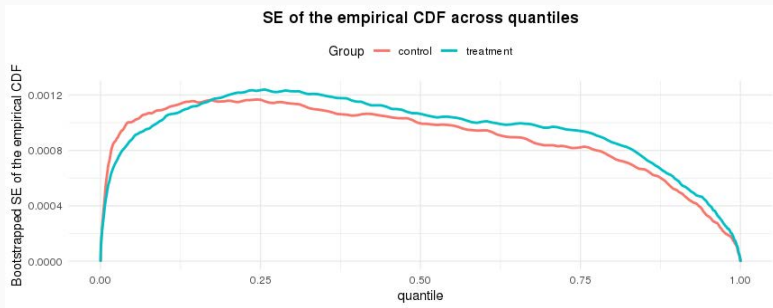
$$\mathbf{V} = \left\{ \underbrace{1, \dots, 1}_{B \text{ repetition}}, \underbrace{2, \dots, 2}_{B \text{ repetition}}, \dots, \underbrace{s, \dots, s}_{B \text{ repetition}} \right\}$$

and then split \mathbf{V} into B vectors of length s , and treat each vector \mathbf{v}_i , $i = 1, \dots, B$ as one bootstrap sample on the buckets. For each sample, count how many $1, \dots, s$ each appears, and weight corresponding $c_{i,j}$ by how many times i -th bucket appears in \mathbf{v}_i , this would give us the empirical cdf at all unique log-scaled values \tilde{X}_j , $j = 1, \dots, K$.

Step 3: calculate bootstrapped SE for empirical CDF

Using all bootstrap samples, estimate the standard deviation of empirical cdf at all \tilde{X}_j (denoted as $\mathbb{F}_{n,i}(\tilde{X}_j)$, $i = 1, \dots, B$) as well, which would approximate v at $\mathbb{F}_n(\tilde{X}_j)$:

$$\begin{aligned}\text{Var}(\mathbb{F}_n(\tilde{X}_j)) &\approx B^{-1} \sum_{i=1}^B \left(\mathbb{F}_{n,i}(\tilde{X}_j) - B^{-1} \sum_{k=1}^B \mathbb{F}_{n,k}(\tilde{X}_k) \right)^2 \\ &:= \hat{v}_n^B[\mathbb{F}_n(\tilde{X}_j)].\end{aligned}$$

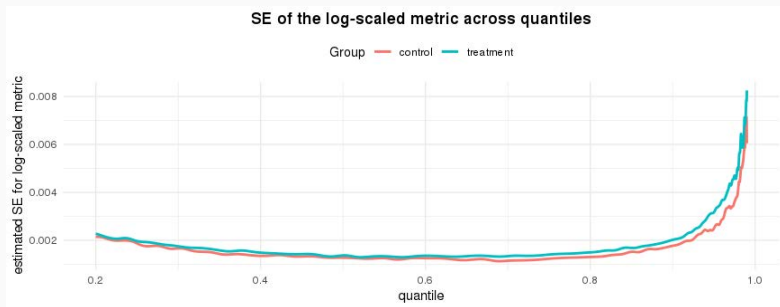


Step 4: appeal to Woodruff CI and convert to SE on quantiles

$$\text{Let } \hat{\xi}_{\mathbb{F}_n(\tilde{X}_j)}^L = \inf \left\{ t : \mathbb{F}_n(t) \geq \mathbb{F}_n(\tilde{X}_j) - \sqrt{\hat{\nu}_n^B[\mathbb{F}_n(\tilde{X}_j)]} \right\},$$

$$\hat{\xi}_{\mathbb{F}_n(\tilde{X}_j)}^U = \inf \left\{ t : \mathbb{F}_n(t) \geq \mathbb{F}_n(\tilde{X}_j) + \sqrt{\hat{\nu}_n^B[\mathbb{F}_n(\tilde{X}_j)]} \right\},$$

$$\text{then let } \widehat{SE}(\tilde{X}_j) = \max \left\{ \tilde{X}_j - \hat{\xi}_{\mathbb{F}_n(\tilde{X}_j)}^L, \hat{\xi}_{\mathbb{F}_n(\tilde{X}_j)}^U - \tilde{X}_j \right\}.$$



Step 5: calculate QTE on a pre-set grid

For a set of interested quantile locations, e.g. 20% to 99% with step size 1%, we can linearly interpolate the variance term for control and treatment using $\widehat{SE}(\tilde{X})$.

QTE:

- $\Delta\%$ at grid:

$$\Delta_i\% = \left(\exp\left(\mathbb{F}_n^{T,-1}(\text{grid}_i)\right) - \exp\left(\mathbb{F}_n^{C,-1}(\text{grid}_i)\right) \right) / \exp\left(\mathbb{F}_n^{C,-1}(\text{grid}_i)\right) \times 100\%;$$

- SE for $\Delta\%$ at grid:

$$\widehat{SE}(\Delta_i\%) = \sqrt{\widehat{SE}\left(\mathbb{F}_n^{C,-1}(\text{grid}_i)\right)^2 + \widehat{SE}\left(\mathbb{F}_n^{T,-1}(\text{grid}_i)\right)^2} \times \frac{\exp\left(\mathbb{F}_n^{T,-1}(\text{grid}_i)\right)}{\exp\left(\mathbb{F}_n^{C,-1}(\text{grid}_i)\right)};$$

- $\Delta\%$ p -value at grid: $p_i = 2 \times \Phi\left(-\left|\frac{\Delta_i\%}{\widehat{SE}(\Delta_i\%)}\right|\right)$;

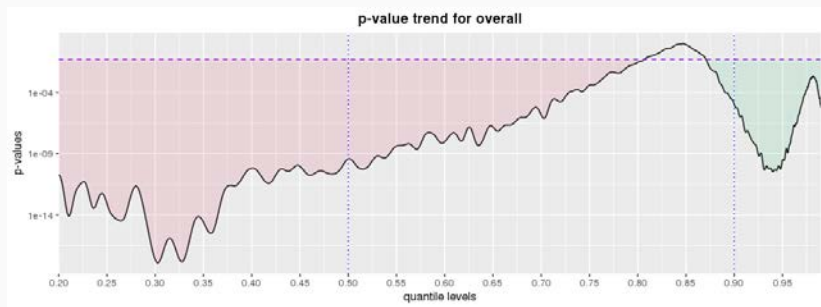
- $\Delta\%$ confidence interval at level $1 - \alpha$ at grid:

$$CI_i = [\Delta_i\% - z_{\alpha/2}\widehat{SE}(\Delta_i\%), \Delta_i\% + z_{\alpha/2}\widehat{SE}(\Delta_i\%)].$$

Evaluation

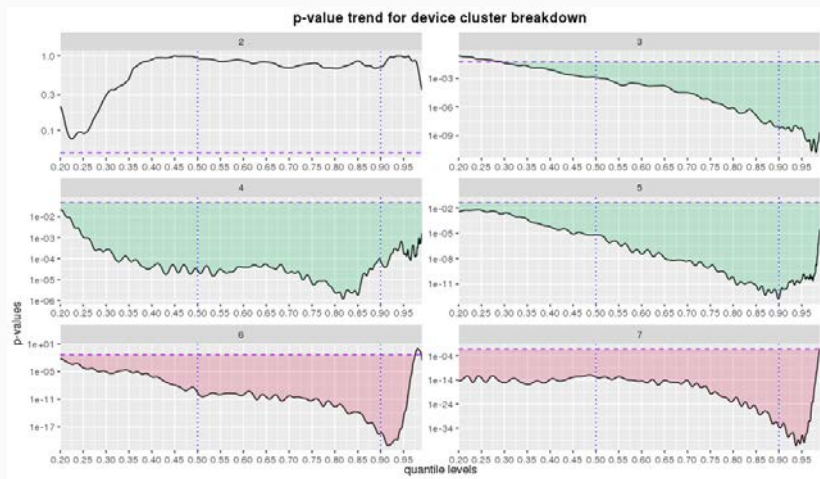
Example: overall

Going back to our motivating example, where P50 and P90 disagrees:

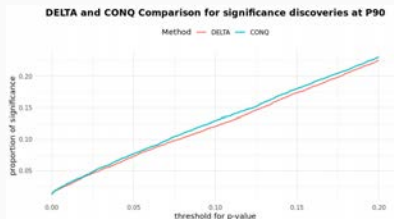
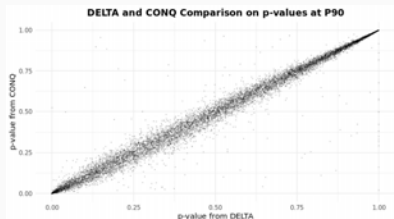
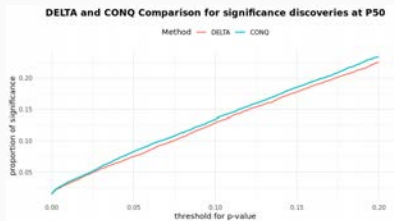
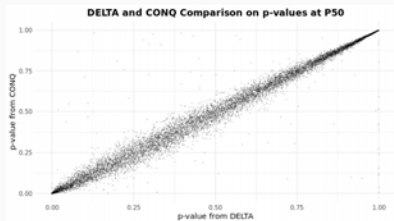


Example: device cluster breakdown

Further breakdown the results by device clusters (low-end to high-end devices), surfacing HTE:



Comparison with the delta-method



A/A tests

Here we pick A/A tests (810 treatment and control pairs) at Snap, and run BH procedure on them at nominal FDR level 0.05, 0.1, 0.2, the number of rejections are shown below:

	0.05	0.1	0.2
P20	0 (0%)	0 (0%)	0 (0%)
P30	0 (0%)	0 (0%)	0 (0%)
P40	0 (0%)	0 (0%)	0 (0%)
P50	0 (0%)	0 (0%)	0 (0%)
P60	0 (0%)	0 (0%)	0 (0%)
P70	0 (0%)	0 (0%)	0 (0%)
P80	0 (0%)	0 (0%)	1 (0.12%)
P90	0 (0%)	0 (0%)	5 (0.62%)
P95	2 (0.25%)	2 (0.25%)	8 (0.99%)

- **Clustering** Users at Snapchat demonstrate distinct in-app engagement activities, how can we cluster them into personas where we can better understand our customers?
- **Causal Inference** What's the causal relationship between app performance metrics to user engagement metrics? e.g., by decreasing my app latency by $X\%$, how much improvement in engagement (say snap sending) can we expect?
- **FDR Control** With so many metrics being tested sequentially and simultaneously, can we efficiently control false positives and false negatives?

Questions?



J. O. Berger.

Statistical decision theory and Bayesian analysis.

Springer Science & Business Media, 2013.



A. Deng, U. Knoblich, and J. Lu.

Applying the delta method in metric analytics: A practical guide with novel ideas.

In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 233–242, 2018.



A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan.

A scalable bootstrap for massive data.

Journal of the Royal Statistical Society: Series B: Statistical Methodology, pages 795–816, 2014.

References ii



M. Liu, X. Sun, M. Varshney, and Y. Xu.

Large-scale online experimentation with quantile metrics.

arXiv preprint arXiv:1903.08762, 2019.



B. W. Silverman.

Density estimation for statistics and data analysis, volume 26.

CRC press, 1986.



R. S. Woodruff.

Confidence intervals for medians and other position measures.

Journal of the American Statistical Association, 47(260):635–646,
1952.



W. B. Wu et al.

On the bahadur representation of sample quantiles for dependent sequences.

The Annals of Statistics, 33(4):1934–1963, 2005.