Why Sampling Your Mobile Analytics Data is Bad For Growth

AMPLITUDE

The costs associated with using high-quality analytics services often force companies into trade-offs of picking which events or users to track, with sampled data being the most common vfate that people resign to.

Unfortunately, sampling can be very harmful, leading to a slew of problems including inaccurate test results and a loss of data integrity. **To make critical product and business decisions based on your data, you need to be able to trust it.**

In this paper, we'll review the case against sampling, including:

- **Acceptable standard error & confidence intervals**
- **The importance of sample size when running A/B tests**
- **Missing out on the long tail**
- **Needing the full set of user data**

Types of Sampling in Analytics

THE PROBLEM: SAMPLING AT THE DATA COLLECTION LEVEL

Before we get into the implications of having all the data available, it's important to understand what sampling really means in behavioral analytics. In the general sense, statistical sampling is the process of measuring a metric on a subset of users in order to estimate the metric across the entire user base. The selection of sampled users can take place either at collection time, i.e. choosing which events to capture at all, or at query time, i.e. choosing which events to analyze (while collecting all of it). The latter is a fine strategy for enabling interactive analysis of large datasets—you don't need 100% accuracy when doing exploratory work, as long as you can choose to get the full results when it matters. **Unfortunately, the way analytics services are priced leads to the former, which means completely losing a sizable chunk of data; this is the problem we want to address.**

Event Level Sampling vs. User Level Sampling

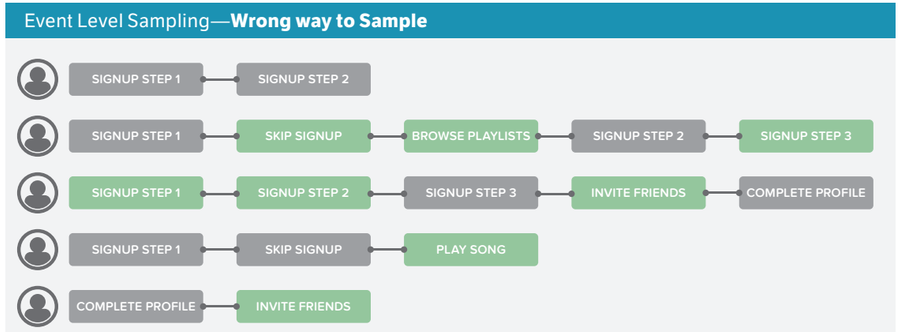
So what does sampling typically look like in our space? It's quite straightforward: choose a subset of users and collect only events performed by those users. A noteworthy but sometimes misunderstood aspect of sampling event data is that it must be performed at the user level (choose whether or not to collect all of a user's events) rather than at the event level (choose whether or not to collect each individual event). Sampling at the user level preserves the calculation of metrics like retention and funnel conversion, which get skewed in non-intuitive ways when you selectively drop events that a user performs. Even getting this part right, however, doesn't guarantee that you can draw accurate conclusions from sampled data.

The Case Against Sampled Data

It's common for power users of analytics to try to get around sampling restrictions by any means possible. Once you hit a volume threshold on Google Analytics, your query results are sampled "in order to reduce latency" and infrastructure costs. People have written an abundance of guides on how to avoid this because they've experienced real-world situations in which sampled query results lead to inaccurate metrics and consequently bad decisions made.

Standard Error and Confidence Intervals

The fundamental trade-off of sampling is, of course, the error in the estimated metric, whether it be retention or a user's lifetime value. The most



In the 2 examples above, we see event timelines from 5 different users. The green events are chosen as part of the sample, while the gray events are not collected. In event level sampling, individual events are sampled for collection. This can result in skewed data and incorrect calculation of important metrics like funnels and retention. User level sampling, in which you collect all events from a select group of users, is the correct way to sample if sampling is required.

relevant aspect of the standard error is that it shrinks with the square root of the sample size.

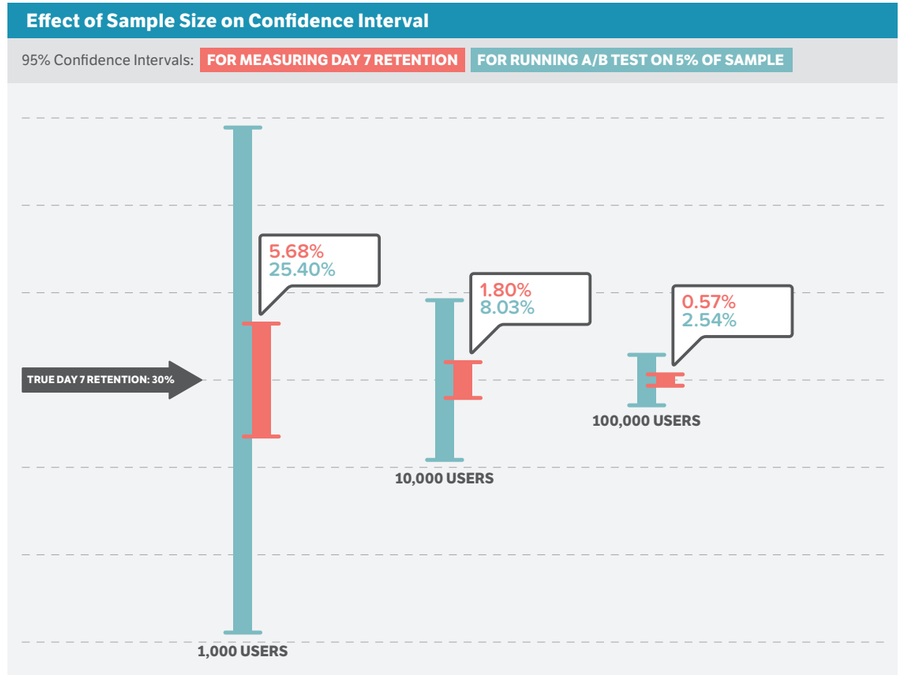
To make this concrete, let's say the true day 7 retention for a cohort of users is 30% and we try to estimate that with a sample. With 10,000 users, the standard error is 0.45% which leads to a potentially acceptable 95% confidence interval of 1.8%. Drop this to 1,000 users, however, and the standard error grows to 1.45% which leads to a very poor 95% confidence interval of 5.68%. You might be thinking that both examples use an unrealistically small number of users; let's say your app has 1,000,000 users, and sampling to 100,000 users gives a confidence interval of 0.57%, so why can't you just do that?

The Importance of Sample Size When Running A/B Tests

One scenario in which that breaks down is while running A/B tests. When experimenting with a significant product or design change, teams will often want to run an A/B test where a small percentage of users are shown the variant and observe how conversion and retention are affected relative to control.

If your data has already been sampled down to 100,000 users, and you show the change to 5% of them, your sample size is significantly cut down and your confidence in the results of the experiment will similarly decrease.

Fast-moving product and growth teams often run tens of A/B tests at once that might each increase conversion by only a few percentage points (see



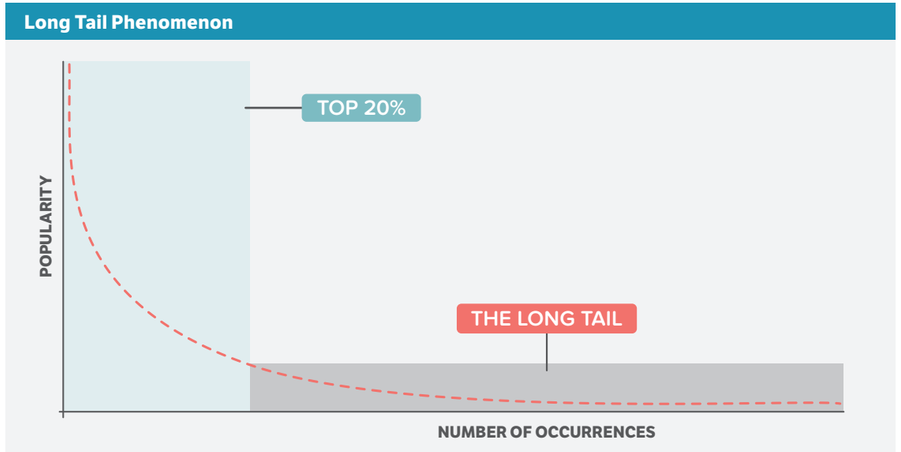
The confidence interval widens as the user sample size decreases. This is particularly important when thinking about confidence intervals for A/B tests. Let's say you collect data from a sample size of 100,000 users. If you want to show 5% of those users a test variant, you will be working with a confidence interval of 2.54%. If you're expecting incremental changes of 2-3% uplift in retention for each test, it will be difficult to draw solid conclusions with a confidence interval of that size.

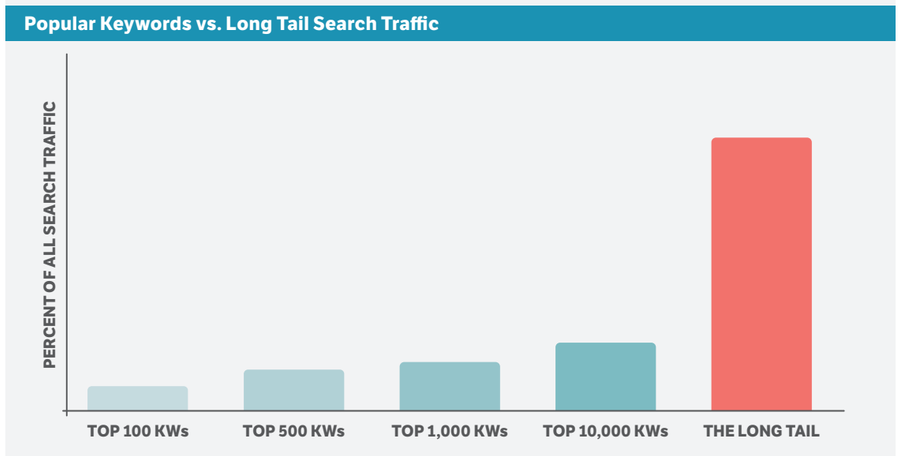
High Tempo Testing, a methodology introduced by Sean Ellis). If you've sampled your user base at collection time too much, you won't be able to draw meaningful conclusions from these tests (or worse, you'll come to the wrong conclusion!). Fortunately, our partner Optimizely has put together some great resources to make sure you don't make this mistake, but you'll need to collect enough data in order to leverage them.

Missing Out on the Long Tail

You've probably heard about the Long Tail phenomenon, which suggests that the true demand curve for users (under certain conditions) has significant area in the "tail," e.g. the lower 80% of "items." Whether it's e-commerce, entertainment, or content, there are many examples of online products and services that realize considerable value outside of the most popular items; it's often considered one of the reasons that they beat out brick and mortar shops that can't compete on those niche items. In order to effectively understand the long tail behavior of users, it's necessary to have enough data. An individual item might only cater to one in a thousand users, so only by capturing all of their events can you hope to perform meaningful analysis on them. For example, it would be completely hopeless to run an A/B test and observe its impact on the long tail if you already sampled a significant portion of your data out.

The long tail in search keywords is a great example of how it can be efficient to focus your attention on not just the most common items, but also those outside of the top ten or twenty percent. It turns out that, while the popular keywords by definition generate the most traffic, they're also very competitive and thus ineffi-





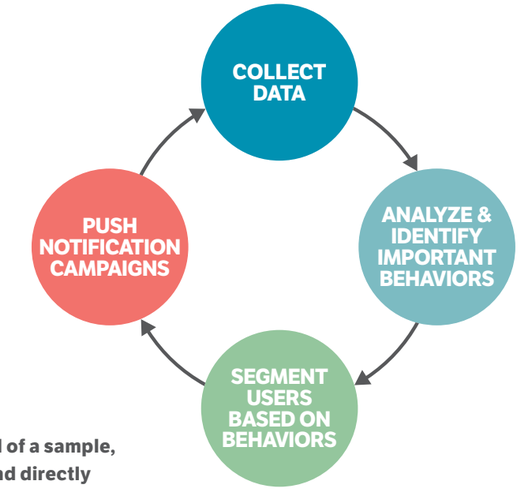
Traffic from long tail keywords makes up 80% of search traffic.
Data from http://neilpatel.com/2015/05/07/a-step-by-step-guide-to-integrating-long-tail-keywords-within-blog-posts/

cient to bid on. Long tail keywords that are more niche to your product or service will be cheaper and yield higher conversion since the user has expressed a more focused desire.

Similarly, in behavioral analytics, the uncommon events performed in an app often reveal the characteristics of a highly-engaged user or whale, which is critical to understanding the features that provide the most value.

These features reveal potential actions that you can optimize and make more prominent in your app to get more users highly engaged and retained.

One of the most powerful features of analytics is the ability to slice and dice a user base by arbitrary properties and analyze a small cohort's behavior in comparison to the total population. Sampling often takes that away because the error becomes unmanageable.



Taking Action

Tracking all of your users, instead of a sample, also enables you to take action and directly communicate with them depending on the behavior they exhibit.

This can range from simple things like sending a friendly welcome email after a user signs up or complex re-engagement strategies based on the last events they performed before churning. We partner with the top mobile marketing automation services to turn insights derived from analytics into concrete actions that improve user engagement, but this is only possible because we have the full list of users and the events they perform. If you only have data for a subset of your users, you can't effectively segment users into campaigns based on their behaviors. (For more on the topic of using behavioral data to inform your mobile engagement strategy, check out this upcoming webinar!)

Another common use case is doing effective customer support. You might not think of an analytics platform as a tool for a support team, but it turns out that knowing the exact sequence of events that a user performed makes debugging issues significantly easier. Being able to look up the device and application information associated with the behavior provides a great starting point for a support engineer to identify or rule out different possibilities.

Since you don't know a priori which users will run into problems, sampling leaves you with a low chance of having the necessary information available to provide the best user experience.

Ultimately, It's About Trusting Your Data

One of the far more insidious but less apparent downsides to sampling is that it adds an additional roadblock to trusting your data. **Knowing that you can trust the data in front of you and make decisions on it is essential to effectively using analytics.**

Sampling puts an extra layer between you and what's going on. Looking at sampled data raises a ton of questions. What if your sampling algorithm made a mistake and what you're looking at is not a representative sample? What if you need to audit the data against another dataset? How do you know that the data is being calculated correctly?

None of these are impossible to work on but they're actually far more destructive to successfully leveraging analytics within an organization.

Moreover, while statistical sampling of data is a common practice for approximating metrics, it is a very limited approach when it comes to analyzing user behavior for growth and engagement. In modern analytics, high-level numbers are no longer enough, and considerable time is spent investigating the long tail of users, running experiments, and re-engaging users through campaigns.

Our goal here at Amplitude is to make sampling an outdated aspect of analytics and empower product & growth managers to make decisions based on data they know they can trust.