18

Variance Estimation and Improved Sensitivity: Pitfalls and Solutions

With great power comes small effect size
– *Unknown*

Why you care*: What is the point of running an experiment if you cannot analyze it in a trustworthy way? Variance is the core of experiment analysis. Almost all the key statistical concepts we have introduced are related to variance, such as statistical significance, p-value, power, and confidence interval. It is imperative to not only correctly estimate variance, but also to understand how to achieve variance reduction to gain sensitivity of the statistical hypothesis tests.*

This chapter covers variance, which is the most critical element for computing p-values and confidence intervals. We primarily focus on two topics: the common pitfalls (and solutions) in variance estimation and the techniques for reducing variance that result in better sensitivity.

Let's review the standard procedure for computing the variance of an average metric, with $i=1, \ldots, n$ independent identically distributed (i.i.d.) samples. In most cases, i is a user, but it can also be a session, a page, a user day, and so on:

- Compute the metric (the average): $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
- Compute the sample variance: $var(Y) = \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- Compute the variance of the average metric which is the sample variance scaled by a factor of n: $var(\bar{Y}) = var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2} * n * var(Y) = \frac{\hat{\sigma}^2}{n}$

Common Pitfalls

If you incorrectly estimate the variance, then the p-value and confidence interval will be incorrect, making your conclusions from the hypothesis test wrong. Overestimated variance leads to false negatives and underestimated variance leads to false positives. Here are a few common pitfalls when it comes to variance estimation.

Delta vs. Delta %

It is very common to use the relative difference instead of the absolute difference when reporting results from an experiment. It is difficult to know if 0.01 more sessions from an average user are a lot or how it compares with the impact on other metrics. Decision makers usually understand the magnitude of a 1% session increase. The relative difference, called *percent delta* is defined as:

$$\Delta\% = \frac{\Delta}{\overline{Y^c}} \tag{18.1}$$

To properly estimate the confidence interval on Δ%, we need to estimate its variance. Variance for the delta is the sum of the variances of each component:

$$var(\Delta) = var\left(\overline{Y^t} - \overline{Y^c}\right) = var\left(\overline{Y^t}\right) + var\left(\overline{Y^c}\right) \tag{18.2}$$

To estimate the variance of Δ%, a common mistake is to divide $var(\Delta)$ by $\overline{Y^c}^2$, that is, $\frac{var(\Delta)}{\overline{Y^c}^2}$. This is incorrect because $\overline{Y^c}$ itself is a random variable. The correct way to estimate the variance is:

$$var(\Delta\%) = var\left(\frac{\overline{Y^t} - \overline{Y^c}}{\overline{Y^c}}\right) = var\left(\frac{\overline{Y^t}}{\overline{Y^c}}\right). \tag{18.3}$$

We will discuss how to estimate the variance of the ratio in the section below.

Ratio Metrics. When Analysis Unit Is Different from Experiment Unit

Many important metrics come from the ratio of two metrics. For example, click-through rate (CTR) is usually defined as the ratio of total clicks to total pageviews; revenue-per-click is defined as the ratio of total revenue to total clicks. Unlike metrics such as clicks-per-user or revenue-per-user, when you use a ratio of two metrics, the analysis unit is no longer a user, but a pageview or click. When the experiment is randomized by the unit of a user, this can create a challenge for estimating variance.

The variance formula $var(Y) = \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is so simple and elegant that it's easy to forget a critical assumption behind it: the samples

($Y_1, \ldots, Y_n$) need to be i.i.d. (independently identically distributed) or at least uncorrelated. This assumption is satisfied if the analysis unit is the same as the experimental (randomization) unit. It is usually violated otherwise. For user-level metrics, each Y_i represents the measurement for a user. The analysis unit matches the experiment unit and hence the i.i.d. assumption is valid. However, for page-level metrics, each Y_i represents a measurement for a page while the experiment is randomized by user, so Y_1, Y_2 and Y_3 could all be from the same user and are "correlated." Because of such "within user correlation," variance computed using the simple formula would be biased.

To correctly estimate the variance, you can write the ratio metric as the ratio of "average of user level metrics," (see Equation 18.4)

$$M = \frac{\bar{X}}{\bar{Y}}. \tag{18.4}$$

Because $\bar{X}$ and $\bar{Y}$ are jointly bivariate normal in the limit, M, as the ratio of the two averages, is also normally distributed. Therefore, by the delta method we can estimate the variance as (Deng et al. 2017) (see Equation 18.5):

$$var(M) = \frac{1}{\bar{Y}^2} var(\bar{X}) + \frac{X^2}{\bar{Y}^4} var(\bar{Y}) - 2\ \frac{\bar{X}}{\bar{Y}^3} cov(\bar{X}, \bar{Y}). \tag{18.5}$$

In the case of Δ%, Y^t and Y^c are independent, hence (see Equation 18.6)

$$var(\varDelta\%) = \frac{1}{\overline{Y^c}^2} var\left(\overline{Y^t}\right) + \frac{\overline{Y^t}^2}{\overline{Y^c}^4} var\left(\overline{Y^c}\right). \tag{18.6}$$

Note that when the Treatment and Control means differ significantly, this is substantially different from the incorrect estimate of $\frac{var(\varDelta)}{\overline{Y^c}^2}$.

Note that there are metrics that cannot be written in the form of the ratio of two user-level metrics, for example, 90th percentile of page load time. For these metrics, we may need to resort to bootstrap method (Efron and Tibshriani 1994) where you simulate randomization by sampling with replacement and estimate the variance from many repeated simulations. Even though bootstrap is computationally expensive, it is a powerful technique, broadly applicable, and a good complement to the delta method.

Outliers

Outliers come in various forms. The most common are those introduced by bots or spam behaviors clicking or performing many pageviews. Outliers have a big impact on both the mean and variance. In statistical testing, the impact on

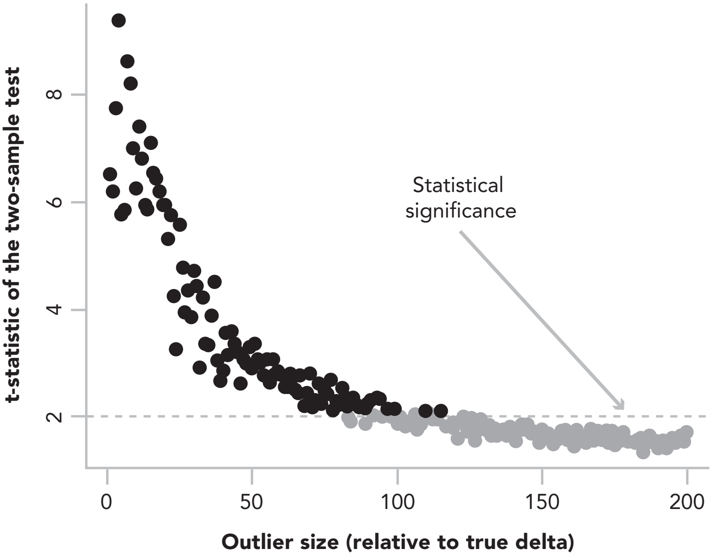

Figure 18.1 In the simulation, as we increase the size of the (single) outlier, the two-sample test goes from being very significant to not significant at all

the variance tends to outweigh the impact on the mean, as we demonstrate using the following simulation.

In the simulation, the Treatment has a positive true delta against Control. We add a single, positive outlier to the Treatment group. The size of the outlier is a multiple of the size of the delta. As we vary the multiplier (the relative size), we notice that while the outlier increases the average of the Treatment, it increases the variance (or the standard deviation) even more. As a result, you can see in Figure 18.1 that the t-statistic decreases as the relative size of the outlier increases and eventually the test is no longer statistically significant.

It is critical to remove outliers when estimating variance. A practical and effective method is to simply cap observations at a reasonable threshold. For example, human users are unlikely to perform a search over 500 times or have over 1,000 pageviews in one day. There are many other outlier removal techniques as well (Hodge and Austin 2004).

Improving Sensitivity

When running a controlled experiment, we want to detect the Treatment effect when it exists. This detection ability is generally referred to as power or sensitivity. One way to improve sensitivity is reducing variance. Here are some of the many ways to achieve a smaller variance:

- Create an evaluation metric with a smaller variance while capturing similar information. For example, the number of searches has a higher variance than the number of searchers; purchase amount (real valued) has higher variance than purchase (Boolean). Kohavi et al. (2009) gives a concrete example where using conversion rate instead of purchasing spend reduced the sample size needed by a factor of 3.3.
- Transform a metric through capping, binarization, or log transformation. For example, instead of using average streaming hour, Netflix uses binary metrics to indicate whether the user streamed more than x hours in a specified time period (Xie and Aurisset 2016). For heavy long-tailed metrics, consider log transformation, especially if interpretability is not a concern. However, there are some metrics, such as revenue, where a log-transformed version may not be the right goal to optimize for the business.
- Use triggered analysis (see Chapter 20). This is a great way to remove noise introduced by people not affected by the Treatment.
- Use stratification, Control-variates or CUPED (Deng et al. 2013). In stratification, you divide the sampling region into strata, sample within each stratum separately, and then combine results from individual strata for the overall estimate, which usually has smaller variance than estimating without stratification. The common strata include platforms (desktop and mobile), browser types (Chrome, Firefox and Edge) and day of week and so on. While stratification is most commonly conducted during the sampling phase (at runtime), it is usually expensive to implement at large scale. Therefore, most applications use post-stratification, which applies stratification retrospectively during the analysis phase. When the sample size is large, this performs like stratified sampling, though it may not reduce variance as well if the sample size is small and variability among samples is big. Control-variates is based on a similar idea, but it uses covariates as regression variables instead of using them to construct the strata. CUPED is an application of these techniques for online experiments, that emphasizes utilization of pre-experiment data (Soriano 2017, Xie and Aurisset 2016, Jackson 2018, Deb et al. 2018). Xie and Aurisset (2016) compare the performance of stratification, post-stratification, and CUPED on Netflix experiments.
- Randomize at a more granular unit. For example, if you care about the page load time metric, you can substantially increase sample size by randomizing per page. You can also randomize per search query to reduce variance if

you're looking at per query metrics. Note that there are disadvantages with a randomization unit smaller than a user:

  - ◦ If the experiment is about making a noticeable change to the UI, giving the same user inconsistent UIs makes it a bad user experience.
  - ◦ It is impossible to measure any user-level impact over time (e.g. user retention).

- Design a paired experiment. If you can show the same user both Treatment and Control in a paired design, you can remove between-user variability and achieve a smaller variance. One popular method for evaluating ranked lists is the interleaving design, where you interleave two ranked lists and present the joint list to user at the same time (Chapelle et al. 2012, Radlinski and Craswell 2013).
- Pool Control groups. If you have several experiments splitting traffic and each has their own Control, consider pooling the separate controls to form a larger, shared Control group. Comparing each Treatment with this shared Control group increases the power for all experiments involved. If you know the sizes of all Treatments you're comparing the Control group with, you can mathematically derive the optimal size for the shared Control. Here are considerations for implementing this in practice:
  - ◦ If each experiment has its own trigger condition, it may be hard to instrument them all on the same Control.
  - ◦ You may want to compare Treatments against each other directly. How much does statistical power matter in such comparisons relative to testing against the Control?
  - ◦ There are benefits of having the same sized Treatment and Control in the comparison, even though the pooled Control is more than likely bigger than the Treatment groups. Balanced variants lead to a faster normality convergence (see Chapter 17) and less potential concern about cache sizes (depending on how you cache implementation).

Variance of Other Statistics

In most discussions in the book, we assume that the statistic of interest is the mean. What if you're interested in other statistics, such as quantiles? When it comes to time-based metrics, such as page-load-time (PLT), it is common to use quantiles, not the mean, to measure site-speed performance. For instance, the 90th or 95th percentiles usually measure user engagement-related load times, while the 99th percentile is more often server-side latency measurements.

While you can always resort to bootstrap for conducting the statistical test by finding the tail probabilities, it gets expensive computationally as data size grows. On the other hand, if the statistic follows a normal distribution asymptotically, you can estimate variance cheaply. For example, the asymptotic variance for quantile metrics is a function of the density (Lehmann and Romano 2005). By estimating density, you can estimate variance.

There is another layer of complication. Most time-based metrics are at the event/page level, while the experiment is randomized at user level. In this case, apply a combination of density estimation and the delta method (Liu et al. 2018).