> *Experiment 2 has the lowest power because of the large variance and small effect*

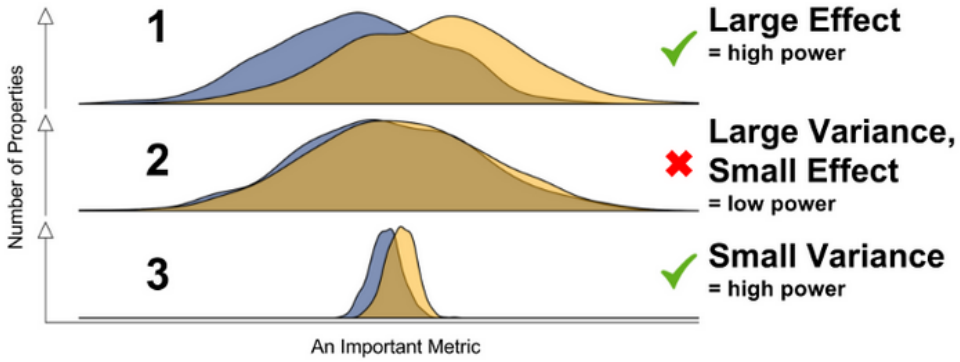

Figure 1. Distributions of properties in base (blue) and variant (yellow) for an important metric in three experiments

CUPED: Controlled-experiment Using Pre-Experiment Data

CUPED is a technique developed by the Experiment Platform team at Microsoft (Deng, Xu, Kohavi, & Walker, 2013) that tries to remove variance in a metric that can be accounted for by pre-experiment information. **Reducing variance helps to increase power to detect small effects.** It's like dealing with the problem in Experiment 2 above, and trying to move to the case of Experiment 3, which has the same effect but with much smaller variance.

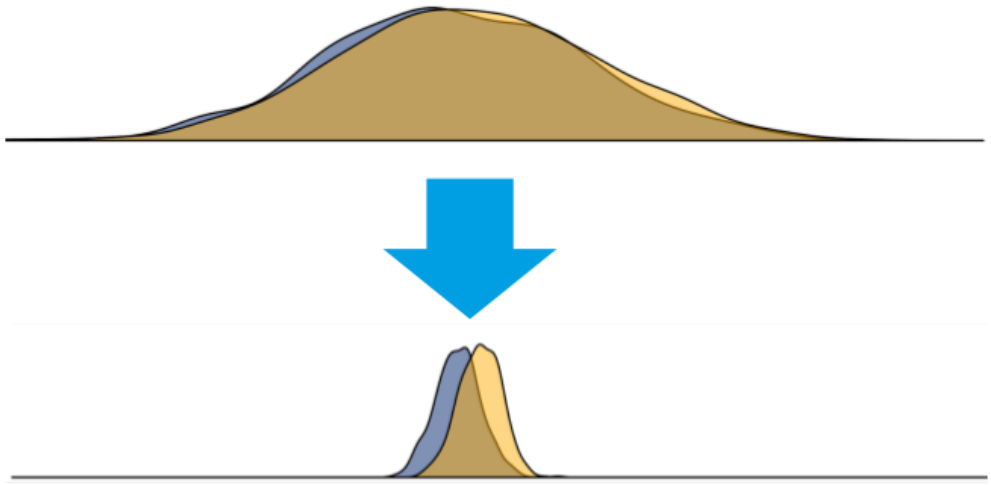

Figure 2. Example of how distribution can change when applying CUPED

Introduction

CUPED (Controlled-experiment Using Pre-Experiment Data) was introduced in Deng et. al. (2013) (Reference 1), and is probably the most used variance reduction technique in A/B testing in the tech industry. At a high level, you can think of CUPED as applying the technique of ***control variates*** (which I've written about in this previous blog post) to the A/B testing set-up. However, CUPED is such an important special case that it is worth describing in full. The paper also contains practical tips on how to implement CUPED and is definite must-read.

Before describing CUPED, let me cover some notation and background. Assume we are in the A/B testing setting and we want to evaluate the impact of some treatment on a response metric. For individual i, let:

- $Y_i(T)$ denote the value of the metric we would see if the individual was given the treatment,
- $Y_i(C)$ denote the value of the metric we would see if the individual was not given the treatment (i.e. was in control),
- Y_i denote the observed value (i.e. $Y_i = Y_i(T)$ or $Y_i = Y_i(C)$, depending on whether i was in treatment or control).

We want to estimate the ***average treatment effect (ATE)*** across individuals, $\Delta = \mathbb{E}[Y_i(T) - Y_i(C)]$. The most commonly used estimator for this is the ***difference-in-means estimator***

$$\begin{aligned}\hat{\Delta} &= \left(\frac{\sum_{i \text{ in treatment}} Y_i}{\#\{i \text{ in treatment}\}}\right) - \left(\frac{\sum_{i \text{ in control}} Y_i}{\#\{i \text{ in control}\}}\right) \\ &=: \overline{Y}_T - \overline{Y}_C.\end{aligned}$$

The difference-in-means estimator is unbiased for the ATE and has a certain variance. CUPED is another estimator for the ATE that is (approximately) unbiased and usually has smaller variance than the difference-in-means estimator.

Описание метода разности разностей удобно сразу осуществить на примере конкретного исследования. Для этого мы воспользуемся работой Карда и Крюгера [*Card*, *Kreuger*, 1994].

В 1992 г. в штате Нью-Джерси, США, минимальный размер оплаты труда был увеличен с 4,25 до 5,05 долл. Экономическая теория подсказывает, что подобное решение должно сказаться на занятости работников с низкой квалификацией (ведь именно их труд часто оплачивается по минимальной ставке). Эту гипотезу решили проверить Кард и Крюгер в своей работе. Они собрали данные о занятости работников в ресторанах быстрого питания. Таким образом, отдельный объект в их выборке — это один ресторан быстрого питания, а зависимая переменная — число работников, занятых в этом ресторане полный рабочий день.

Таким образом, средний эффект воздействия (*average treatment effect*), который интересует авторов работы, — это среднее изменение занятости в ресторане быстрого питания в Нью-Джерси в результате принятия нового закона о минимальной заработной плате.

Как можно было бы подсчитать это изменение?

Первый, довольно наивный, подход — взять данные по Нью-Джерси до и после изменения минимальной заработной платы и сравнить их между собой. Этот подход плох, потому что занятость с течением времени может изменяться не только в результате изменения заработной платы, но и по каким-то другим причинам. Есть множество глобальных факторов, которые влияют на всю Америку в целом и могли сказаться на занятости. Сравнивая такие средние значения до и после, мы не можем понять, влияет ли на происходящее минимальная зарплата или дело в каких-то других факторах. Например, в США могла начаться рецессия, которая способствовала бы снижению занятости во всех штатах независимо от экономической политики на рынке труда. Или в ресторанном бизнесе могла быть внедрена новая продвинутая технология, которая позволила снизить спрос на труд низкоквалифицированных работников в этой отрасли.

Второй подход состоит в том, чтобы сравнить занятость в среднем ресторане в Нью-Джерси (т.е. в испытуемой группе) со средней занятостью в каком-нибудь другом штате, где минимальная зарплата не изменилась (т.е. в контрольной группе), например в Пенсильвании, в которой в тот же самый период времени минимальная заработная плата осталась на прежнем уровне. Понятно, что такой подход тоже не совершенен. В отличие от совсем «честного» эксперимента, когда мы случайным образом делим рестораны на две группы, здесь эксперимент не совсем чистый, потому что все рестораны первой группы находятся в одном штате, а рестораны второй группы — в другом. И эти штаты, вполне возможно, хотя они и похожи, отличаются не только минимальной зарплатой, но какими-то еще характеристиками. И поэтому снова невозможно выяснить, объясняются ли различия в занятости в этих двух штатах именно различием в минимальной заработной плате или чем-то еще.

Вместо каждого из двух указанных подходов можно применить альтернативный метод, который объединяет их преимущества и помогает избежать недостатков. Чтобы понять, как он работает, перечислим основные факторы, которые могут влиять на занятость в типичном ресторане:

- специфические особенности штата, в котором расположен ресторан (эффект штата);
- особенности различных периодов времени, скажем, изменение экономической конъюнктуры (временной эффект);
- эффект изменения минимальной заработной платы (тот самый эффект, который мы пытаемся оценить).

Формально мы можем записать это так:

$$Y_{ist} = \alpha_s + \mu_t + \delta \cdot D_{ist} + \varepsilon_{ist},$$

где индекс i — номер ресторана; индекс s — штат (Нью-Джерси или Пенсильвания); индекс t — момент времени (период **до** изменения заработной платы в Нью-Джерси или период **после** него);

Геометрическая интерпретация применения метода разности разностей представлена на рис. 11.1. Пунктирная линия на нем показывает, как менялась бы занятость в Нью-Джерси, если бы этот штат не подвергся воздействию. Подход разности разностей предполагает, что в этом случае динамика занятости была бы аналогична контрольной группе (что в нашем примере означает снижение числа работников).

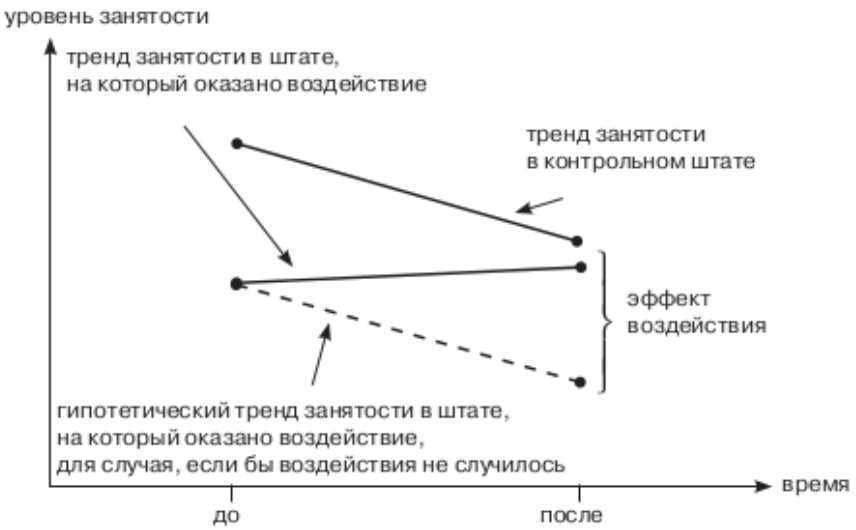

Рис. 11.1. Геометрическая иллюстрация метода разности разностей

Метод разности разностей непосредственно связан с оценкой моделей при помощи регрессий. В частности, если вы располагаете панельными данными об объектах из испытуемой и контрольной групп за два периода (до и после проведения политики), оценка метода разности разностей может быть получена в результате применения следующей модели:

$$Y_{it} = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot z_t + \delta \cdot x_i \cdot z_t + \varepsilon_{it}, \qquad (11.3)$$

где x_i — бинарная переменная, которая равна единице, если i-й ресторан расположен в Нью-Джерси (т.е. относится к испытуемой группе);

z_t — бинарная переменная, которая равна единице для всех наблюдений, относящихся ко второму периоду (периоду после повышения минимальной зарплаты).

В этом случае, применив МНК, получим следующее уравнение:

$$\hat{Y}_{it} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i + \hat{\beta}_2 \cdot z_t + \hat{\delta} \cdot x_i \cdot z_t.$$

Можно доказать (см. задание в конце главы), что:

$$\hat{\beta}_0 = \bar{Y}_{control,before};$$

$$\hat{\beta}_1 = \bar{Y}_{treatment,before} - \bar{Y}_{control,before};$$

$$\hat{\beta}_2 = \bar{Y}_{control,after} - \bar{Y}_{control,before};$$

$$\hat{\delta} = [\bar{Y}_{treatment,after} - \bar{Y}_{treatment,before}] - [\bar{Y}_{control,after} - \bar{Y}_{control,before}].$$

Таким образом, коэффициент при произведении $x_i \cdot z_i$ равен той самой оценке эффекта воздействия, которую мы вывели выше.

Эквивалентный способ оценивания состоит в применении МНК к следующей парной регрессии:

$$\Delta Y_i = \beta_2 + \delta \cdot x_i + u_i, \qquad (11.4)$$

где x_i — по-прежнему бинарная переменная, которая равна единице, если i-й объект относится к испытуемой группе; $\Delta Y_i = Y_{i1} - Y_{i0}$.

Если применить к этой модели МНК, то оценка коэффициента при регрессоре снова будет задаваться формулой (11.2).

Для состоятельности этой оценки требуется, чтобы в уравнении (11.3) выполнялись все предпосылки модели со стохастическим регрессором (см. гл. 6). В частности, требуется экзогенность объясняющей переменной. Если регрессор оказывается эндогенным из-за пропуска других существенных факторов, влияющих на ΔY, то эта проблема может быть, как обычно, решена включением в уравнение (11.3) контрольных переменных.

Иными словами, метод разности разностей может быть дополнен при помощи учета контрольных переменных, что делает его взаимосвязь с прочими моделями на панельных данных еще более тесной.

В то же время у метода разности разностей есть важное преимущество по сравнению с ними. Дело в том, что иногда вместо панельных данных вам доступны лишь так называемые повторяющиеся пространственные данные. Это означает, что для разных периодов имеются данные по различным объектам. Скажем, в нашем примере про рестораны быстрого питания это означало бы, что до изменения заработной платы исследователи опросили бы одни рестораны, а после изменения — другие. Тогда непосредственно оценить уравнение (11.3) было бы невозможно, так как нельзя было бы рассчитать разности $\Delta Y_i = Y_{i1} - Y_{i0}$ (потому что по каждому отдельному объекту доступны данные только для одного из двух периодов). Однако интересующий нас эффект воздействия все еще мог бы быть рассчитан при помощи метода разности разностей по формуле (11.2).

Метод разности разностей широко используется в современных прикладных исследованиях для анализа последствий применения тех или иных мер экономической политики.

Y_{ist} — число работников, занятых в данном ресторане;

переменная D равна единице в ресторанах, которые находились в Нью-Джерси в тот период, когда там поменялась заработная плата, и равна нулю во всех остальных случаях;

α_s — эффект штата. Он имеет два значения: $\alpha_{control}$, если наблюдение относится к контрольной группе, т.е. к Пенсильвании; $\alpha_{treatment}$, если наблюдение относится к испытуемой группе, т.е. к Нью-Джерси;

μ_t — временной эффект. Он равен μ_{before} до изменения заработной платы и μ_{after} после изменения;

δ — эффект воздействия увеличения заработной платы на занятость. Это тот самый эффект, который требуется оценить;

ε_{ist} — случайные ошибки модели.

Определим ожидаемое количество занятых в ресторане Нью-Джерси **до** изменения заработной платы:

$$E(Y_{ist}|s = treatment,\ t = before) = \mu_{before} + \alpha_{treatment}.$$

Определим ожидаемое количество занятых в ресторане Нью-Джерси **после** изменения:

$$E(Y_{ist}|s = treatment,\ t = after) = \mu_{after} + \alpha_{treatment} + \delta.$$

Вычитая из второго математического ожидания первое, получим ожидаемое изменение занятости в Нью-Джерси:

$$\Delta_{treatment} = \mu_{after} - \mu_{before} + \delta.$$

Теперь определим ожидаемое количество занятых в ресторане Пенсильвании до изменения заработной платы:

$$E(Y_{ist}|s = control,\ t = before) = \mu_{before} + \alpha_{control}.$$

Определим ожидаемое количество занятых в ресторане Пенсильвании после изменения:

$$E(Y_{ist}|s = control,\ t = after) = \mu_{after} + \alpha_{control}.$$

Аналогично прошлому случаю, вычитая из второго математического ожидания первое, получим ожидаемое изменение занятости в Пенсильвании:

$$\Delta_{control} = \mu_{after} - \mu_{before}.$$

Осталось вычесть из первой разности ($\Delta_{treatment}$) вторую разность ($\Delta_{control}$), т.е. найти ту самую разность разностей, которая дала название анализируемому методу:

$$\Delta_{treatment} - \Delta_{control} = \delta.$$

Таким образом, мы показали, что интересующий нас эффект воздействия может быть представлен как разность разностей условных математических ожиданий:

$$\delta = \Delta_{treatment} - \Delta_{control} =$$

$$= [E(Y_{ist} | s = treatment,\ t = after) - E(Y_{ist} | s = treatment, t = before)] -$$

$$-[E(Y_{ist} | s = control,\ t = after) - E(Y_{ist} | s = control,\ t = before)].$$

В силу закона больших чисел состоятельной оценкой каждого из этих математических ожиданий является соответствующее среднее значение. Следовательно, состоятельная оценка эффекта воздействия в данном случае может быть рассчитана так:

$$\hat{\delta} = [\bar{Y}_{treatment,\ after} - \bar{Y}_{treatment,\ before}] - [\bar{Y}_{control,\ after} - \bar{Y}_{control,\ before}], \quad (11.2)$$

где $\bar{Y}_{control,before}$ — средний уровень зависимой переменной в контрольной группе до осуществления воздействия (в нашем случае — это средний уровень занятости в ресторанах Пенсильвании до изменения минимальной заработной платы в Нью-Джерси);

$\bar{Y}_{control,after}$ — средний уровень зависимой переменной в контрольной группе после осуществления воздействия (в нашем случае — это средний уровень занятости в ресторанах Пенсильвании после изменения минимальной заработной платы в Нью-Джерси);

$\bar{Y}_{treatment,before}$ — средний уровень зависимой переменной в испытуемой группе до осуществления воздействия (в нашем случае — это средний уровень занятости в ресторанах Нью-Джерси до изменения минимальной заработной платы);

$\bar{Y}_{treatment,after}$ — средний уровень зависимой переменной в испытуемой группе после осуществления воздействия (в нашем случае — это средний уровень занятости в ресторанах Нью-Джерси после изменения минимальной заработной платы).

What really amazed me was the popularity of the algorithm in the industry. CUPED was first introduced by Microsoft researchers Deng, Xu, Kohavi, Walker (2013) and has been **widely used in companies** such as Netflix, Booking, Meta, Uber, Airbnb, Linkedin, TripAdvisor, DoorDash, Faire, and many others. While digging deeper into it, I noticed a similarity with some causal inference methods I was familiar with, such as Difference-in-Differences or regression with control variables. I was curious and decided to dig deeper.

controlled experiments utilizing pre-experiment data

1. Возьмем случайную переменную X, независимую от Y
2. Представим новую метрику как разность Y и Тета X
3. Необходимо определить Тета

$$Y_{CUPED} = Y - \theta X$$

4. Дисперсия вычисляется по формуле:

$$(\mathrm{var}(Y) + \theta^2 \mathrm{var}(X) - 2\theta \mathrm{cov}(Y, X))$$

1. Дисперсия минимизируется когда:

$$\theta = \mathrm{cov}(Y, X)/\mathrm{var}(X)$$

2. Итоговая дисперсия

$$var_{srs}(Y_{CUPED})_{min} = var_{srs}(Y)(1 - \rho^2)$$

- Дисперсия обусловлена двумя факторами

1. Зачастую нас интересует разница средних
2. Для каждой группы необходимо использовать единую Тета
3. Логичным выбором метрик X будет значение Y на предшествующее эксперименту

When the metrics are mean-centered, another way to look at what is going on is a scatter plot of our metric (y-axis) against our pre-experiment covariate (x-axis). The slope of the line that best fits these points is defined by `theta`, and the simplified CUPED-adjusted score is the vertical distance from each point to the line.

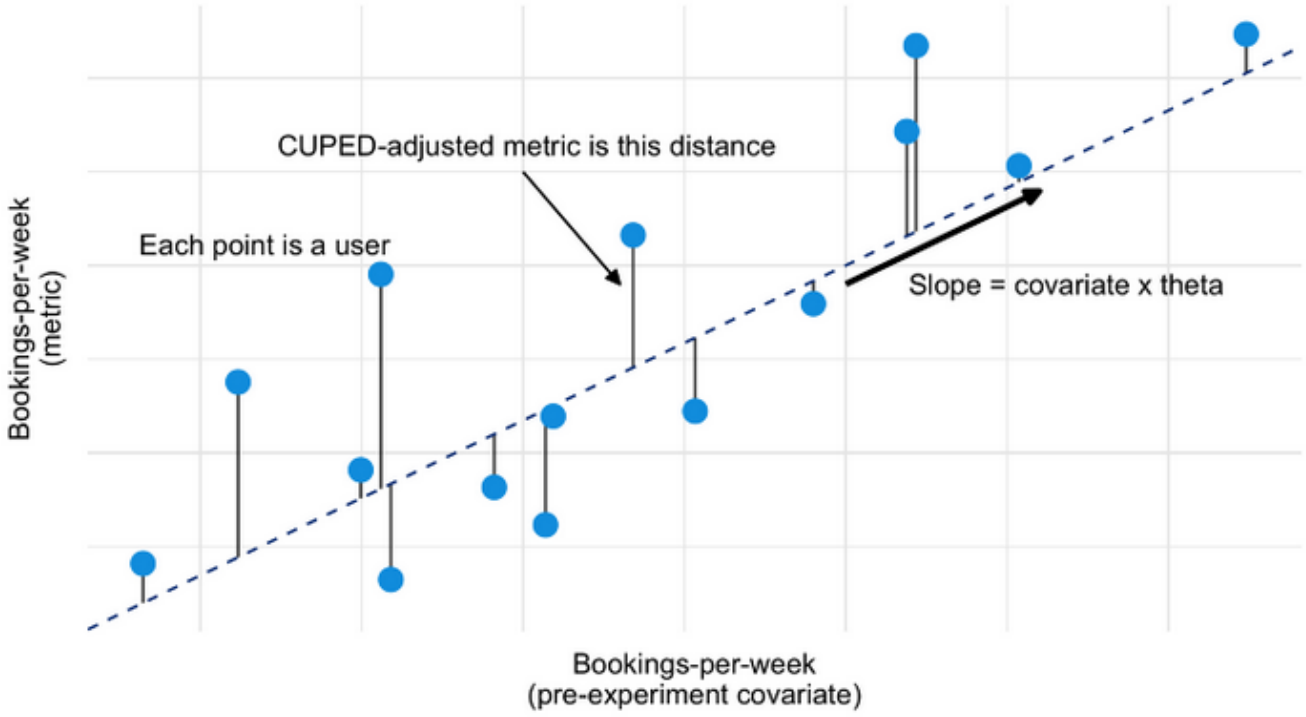

Figure 3. A visual example of how to compute a CUPED-adjusted metric for mean-centered metrics

Statisticians will recognize that `theta` is equivalent to the unstandardized coefficient from an ordinary-least-squares regression of the metric on the pre-experiment covariate. The practical difference is in the computation of this coefficient and the lack of an intercept term, which is not required for variance reduction.

Key idea

Let's focus on just estimating $\mathbb{E}[Y_i(T)]$. The difference-of-means estimator estimates this with $\overline{Y}_T$. Imagine that on top of collecting metric values $Y_1, Y_2, \ldots, Y_{n_t}$ in the treatment group, we also collected pre-experiment values on another (real-valued) variable $X_1, X_2, \ldots, X_{n_t}$. Let's also assume that we know the mean of X (which denote by $\mathbb{E}[X]$). For any fixed parameter θ, we have

$$\begin{aligned}\mathbb{E}[Y_i(T)] &= \mathbb{E}[\overline{Y}_T] \\ &= \mathbb{E}[\overline{Y}_T - \theta X] + \theta\mathbb{E}[X] \\ &= \mathbb{E}[\overline{Y}_T - \theta\overline{X}_T] + \theta\mathbb{E}[X].\end{aligned}$$

Hence,

$$\widetilde{Y}_T = \overline{Y}_T - \theta\overline{X}_T + \theta\mathbb{E}[X]$$

is an unbiased estimator for $\mathbb{E}[Y_i(T)]$. Working through some variance computations, we can show that the variance of $\widetilde{Y}_T$ is minimized when $\theta = \mathrm{Cov}(Y, X)/\mathrm{Var}(X)$, and at this value of θ, we have

$$\mathrm{Var}(\widetilde{Y}_T) = (1 - \rho^2)\mathrm{Var}(\overline{Y}_T) \leq \mathrm{Var}(\overline{Y}_T),$$

where ρ is the correlation between Y and X.

Dotting our i's and crossing our t's

Before we can use $\widetilde{Y}_T = \overline{Y}_T - \theta\overline{X}_T + \theta\mathbb{E}[X]$ as an estimator, we need to address 3 issues.

First, we don't know the value of $\theta = \text{Cov}(Y, X)/\text{Var}(X)$. Notice that θ is simply the population regression coefficient for X when we regress Y on X. Hence, we can replace θ with its sample quantity $\hat{\theta}$, the regression coefficient for Y on X with the sample that we have: $(X_1, Y_1), \ldots, (X_{n_t}, Y_{n_t})$.

This approximation causes $\widetilde{Y}_T$ to no longer be exactly unbiased, because $\mathbb{E}[\hat{\theta}\overline{X}_T] \neq \theta\mathbb{E}[X]$ in general: both $\hat{\theta}$ and $\overline{X}_T$ depend on $X_1, \ldots, X_{n_t}$, complicating the expectation computation. If we want exact unbiasedness, we can use a subsample to estimate $\hat{\theta}$, then use the rest of the sample in the expression for $\widetilde{Y}_T$. (It's usually not worth the effort to do so.)

Second, we don't know the value of $\mathbb{E}[X]$. We can't simply use the sample mean as an estimate for it, because plugging that in simply reduces $\widetilde{Y}_T$ to the original sample mean $\overline{Y}_T$. We could use a subsample to estimate $\mathbb{E}[X]$, then use the rest of the sample in the expression for $\widetilde{Y}_T$.

In the A/B testing setting, we don't have to do anything that fancy! Remember that the quantity we are really interested in is not $\mathbb{E}[Y_i(T)]$ but $\Delta = \mathbb{E}[Y_i(T)] - \mathbb{E}[Y_i(C)]$. Using analagous reasoning for estimating $\mathbb{E}[Y_i(C)]$, we see that

$$\begin{aligned}\widetilde{Y}_T - \widetilde{Y}_C &= (\overline{Y}_T - \theta\overline{X}_T + \theta\mathbb{E}[X]) - (\overline{Y}_C - \theta\overline{X}_C + \theta\mathbb{E}[X]) \\ &= (\overline{Y}_T - \theta\overline{X}_T) - (\overline{Y}_C - \theta\overline{X}_C)\end{aligned}$$

is an unbiased estimator for Δ as well. The $\theta\mathbb{E}[X]$ cancels out, so we don't have to estimate it.

Third, we don't know which X to use. In theory, we can use any variable X. However, recall the variance computation

$$\text{Var}(\widetilde{Y}_T) = (1 - \rho^2)\text{Var}(\overline{Y}_T) \leq \text{Var}(\overline{Y}_T),$$

where ρ is the correlation between Y and X. Thus, we want to pick variables that are most correlated with the metric that we are measuring. Deng et. al. (2013) note that ***in the A/B testing setting, the same metric we want to estimate (Y) but evaluated on a pre-experiment time period often gives the most variance reduction***. This often makes sense: e.g. for engagement metrics, users who are highly engaged before the experiment tend to be highly engaged during the experiment as well.

An important caution here is that X must not be affected by the experiment's treatment. (All pre-experiment variables meet this requirement, Reference 1 adds some other possibilities in Section 4.3.) This is because for CUPED to be unbiased, we assumed that $\mathbb{E}[X]$ has the same value for the treatment and control populations. If X is affected by the treatment such that $\mathbb{E}[X]$ differs across the treatment arms, CUPED will be biased.

Some other notes

- There is an obvious generalization to go from one control variate X to multiple control variates $X_1, \ldots, X_K$: see my previous blog post for some details.
- There is a strong connection between ***stratification*** and CUPED: see Section 3.3 and Appendix A of Reference 1.
- The discussion above applies to the estimation of the treatment effect for ***count metrics*** and not to ratio metrics (see this previous post for definitions of count and ratio metrics). See Appendix B of Reference 1 for how to apply CUPED to ratio metrics.

Recommendations

Let me end this post off with the 4 CUPED recommendations listed in the paper (emphasis mine):

1. **Variance reduction works best for metrics where the distribution varies significantly across the user population.** One common class of such metrics [is] where the value is very different for light and heavy users. Queries-per-user is a paradigmatic example of such a metric.
2. **Using the metric measured in the pre-period as the covariate typically provides the best variance reduction.**
3. **Using a pre-experiment period of 1-2 weeks works well for variance reduction.** Too short a period will lead to poor matching, whereas too long a period will reduce correlation with the outcome metric during the experiment period.
4. **Never use covariates that could be affected by the treatment, as this could bias the results.** We have shown an example where directionally opposite conclusions could result if this requirement is violated.

Frisch-Waugh-Lovell and VaReSE

Ragnar Frisch and Frederick Waugh published the relevant result in the fourth issue of *Econometrica* back in 1933. Incidentally, Frisch, an economist from Norway who also happened to train as a silversmith, coined the term *econometrics* and later won the first Nobel Prize in Economics. So maybe it's worth paying attention to. Michael Lovell extended Frisch and Waugh's result in the 1960s, plastering his name on the now-famous Frisch-Waugh-Lovell theorem, and more recently published a simple proof of it here.

What do we do with the Frisch-Waugh-Lovell theorem? Well, a couple of things. First, *it's not necessary to perform CUPED in two steps.* A statistically identical hypothesis test can be performed just by including the treatment indicator in the first regression and testing it for significance.[3] I'll reproduce the combined specification just so you have it handy:

$$Y = \mu + Z\beta + T \times 1_{treated} + \epsilon$$

Unless you're dealing with massive matrices or something, there is no reason to split the analysis in two.

The second implication is where things get a little more interesting. By reformulating CUPED as a regression with covariates and a treatment effect, we can push the variance reduction further than the CUPED authors advertised as being possible.

Remember the side-tour above about treatment effects (between-group variance) contributing to total variance? Any experiment, by definition, is a potential source of variance in the outcome. (That's the whole point of running experiments!) Using this linear-regression framework, we can modify the combined equation to include not just a *single* assignment variable, but the full *vector* of treatment assignments from all of the K experiments that are running simultaneously:

$$Y = \mu + Z\beta + T_1 + T_2 + \ldots + T_K + \epsilon$$

(Indicator variables omitted for clarity.) As long as the assignments are uncorrelated, that is, properly randomized, the estimates of the treatment effects will remain unbiased, but the standard errors will be smaller than they were with separate regressions.

(Indicator variables omitted for clarity.) As long as the assignments are uncorrelated, that is, properly randomized, the estimates of the treatment effects will remain unbiased, but the standard errors will be smaller than they were with separate regressions.

A linear regression with a treatment vector and CUPED covariates neatly decomposes the outcome variance into three parts: the variance introduced by the experiments, the variance introduced by the covariates, and the residual or unexplained variance. Including effective ($T \neq 0$) experiments into the treatment vector necessarily reduces the residual variance, and therefore produces larger t-statistics on the entire vector of treatment effects. Cue the singing of the statistics angels, and a fresh infusion of color into the experiment dashboard.

The usual advice with running concurrent experiments is "don't worry, the other treatment effects will balance each other out." This is true insofar with regard to the *estimates*, but it's not true with regard to the *variance*, which is additive. So if you're running experiments concurrently, but analyzing them in isolation, you're missing an opportunity to pull out the outcome variance that each experiment is introducing into all the contemporaneous experiments.

If you need a nifty Italianate acronym for the above method, may I suggest VaReSE: Variance Reduction in Simultaneous Experiments.

What's the potential reduction in variance and therefore future sample sizes with VaReSE? In light of the law of total variance, the percent reduction will equal the sum of all the between-group variances for a given round of experiments divided by the residual variance from CUPED. To make an equation out of the preceding sentence:

$$\Delta_{VaReSE} = \frac{\sum_k \frac{1}{4} T_k^2}{\sigma^2_{CUPED}}$$

Depending on the maturity of the experimentation operation, a sample-size reduction of a few percent seems achievable. Notice that this technique benefits from including experiments with negative as well as positive effects; as long as experiments are a source of variance, multiple regression can account for that additional noise.

As a final implementation note, you'll probably want to flip on White standard errors in order to account for treatments whose within-group variance differs from the control, the same as Welch's t-test does.

3.1 Stratification

Stratification is a common technique used in Monte Carlo sampling to achieve variance reduction. In this section, we show how it can be adapted to achieve the same goal in the world of online experimentation.

3.1.1 *Stratification in Simulation*

The basic idea of stratification is to divide the sampling region into strata, sample within each stratum separately and then combine results from individual strata together to give an overall estimate, which usually has a smaller variance than the estimate without stratification.

Mathematically, we want to estimate $\mathbb{E}(Y)$, the expected value of Y, where Y is the variable of interest. The standard Monte Carlo approach is to first simulate n independent samples $Y_i, i = 1, \ldots, n$, and then use the sample average $\overline{Y}$ as the estimator of $\mathbb{E}(Y)$. $\overline{Y}$ is unbiased and $\text{var}(\overline{Y}) = \text{var}(Y)/n$.

Let's consider a more strategic sampling scheme. Assume we can divide the sampling region of Y into K subregions (strata) with w_k the probability that Y falls into the kth stratum, $k = 1, \ldots, K$. If we fix the number of points sampled from the kth stratum to be $n_k = n \cdot w_k$, we can define a stratified average to be

$$\widehat{Y}_{strat} = \sum_{k=1}^{K} w_k \overline{Y}_k, \tag{2}$$

where $\overline{Y}_k$ is the average within the kth stratum.

The stratified average $\widehat{Y}_{strat}$ and the standard average $\overline{Y}$ have the same expected value but the former gives a smaller variance when the means are different across the strata. The intuition is that the variance of $\overline{Y}$ can be decomposed into the within-strata variance and the between-strata variance, and the latter is removed through stratification. For example, the variance of children's heights in general is large. However, if we stratify them by their age, we can get a much smaller variance within each age group. More formally,

$$\begin{aligned}\text{var}(\overline{Y}) &= \sum_{k=1}^{K} \frac{w_k}{n}\sigma_k^2 + \sum_{k=1}^{K} \frac{w_k}{n}(\mu_k - \mu)^2 \\ &\geq \sum_{k=1}^{K} \frac{w_k}{n}\sigma_k^2 = \text{var}(\widehat{Y}_{strat})\end{aligned}$$

where (μ_k, σ_k^2) denote the mean and variance for users in the kth stratum. More detailed proof can be found in standard Monte Carlo books (e.g. Asmussen and Glynn (2008)). A good stratification is the one that aligns well with the underlying clusters in the data. By explicitly identifying these clusters as strata, we essentially remove the extra variance introduced by them.

3.1.2 *Stratification in Online Experimentation*

In the online world, because we collect data as they arrive over time, we are usually unable to sample from strata formed ahead of time. However, we can still utilize pre-experiment variables to construct strata after all the data are collected (for theoretical justification see Asmussen and Glynn (2008, Page 153)). For example, if Y_i is the number of queries from a user i, a covariate X_i could be the browser that the user used before the experiment started. The stratified average in (2) can then be computed by grouping Y according to the value of X,

$$\widehat{Y}_{strat} = \sum_{k=1}^{K} w_k \overline{Y}_k = \sum_{k=1}^{K} w_k \left(\frac{1}{n_k} \sum_{i:X_i=k} Y_i \right).$$

Using superscripts to denote treatment and control groups, the stratified delta

$$\Delta_{strat} = \widehat{Y}^{(t)}_{strat} - \widehat{Y}^{(c)}_{strat} = \sum_{k=1}^{K} w_k (\overline{Y}^{(t)}_k - \overline{Y}^{(c)}_k)$$

enjoys the same variance reduction as the stratified average in Eq. (2). It is important to note that by using only the pre-experiment information, the stratification variable X is independent of the experiment effect. This ensures that the stratified delta is unbiased.

In practice, we don't always know the appropriate weights w_k to use. In the context of online experimentation, these can usually be computed from users not in the experiment. As we will see in Section 3.3, when we formulate the same problem in the form of control variates (Section 3.2), we no longer need to estimate the weights.

3.2.1 Control Variates in Simulation

The idea of variance reduction through control variates stems from the following observation. Assume we can simulate another random variable X in addition to Y with known expectation $\mathbb{E}(X)$. In other words, we have independent pairs of $(Y_i, X_i), i = 1, \ldots, n$. Define

$$\widehat{Y}_{cv} = \overline{Y} - \theta\overline{X} + \theta\mathbb{E}X, \tag{3}$$

where θ is any constant. $\widehat{Y}_{cv}$ is an unbiased estimator of $\mathbb{E}(Y)$ since $-\theta\mathbb{E}(\overline{X}) + \theta\mathbb{E}(X) = 0$. The variance of $\widehat{Y}_{cv}$ is

$$\begin{aligned}\text{var}(\widehat{Y}_{cv}) &= \text{var}(\overline{Y} - \theta\overline{X}) = \text{var}(Y - \theta X)/n \\ &= \frac{1}{n}(\text{var}(Y) + \theta^2\text{var}(X) - 2\theta\text{cov}(Y, X)).\end{aligned}$$

Note that $\text{var}(\widehat{Y}_{cv})$ is minimized when we choose

$$\theta = \text{cov}(Y, X)/\text{var}(X) \tag{4}$$

and with this optimal choice of θ, we have

$$\text{var}(\widehat{Y}_{cv}) = \text{var}(\overline{Y})(1 - \rho^2), \tag{5}$$

where $\rho = \text{cor}(Y, X)$ is the correlation between Y and X. Compare (5) to the variance of $\overline{Y}$, the variance is reduced by a factor of ρ^2. The larger ρ, the better the variance reduction. The single control variate case can be easily generalized to include multiple variables.

It is interesting to point out the connection with linear regression. The optimal $\boldsymbol{\theta}$ turns out to be the ordinary least square (OLS) solution of regressing (centered) Y on (centered) X, which gives variance

$$\text{var}(\widehat{Y}_{cv}) = \text{var}(\overline{Y})(1 - R^2),$$

with R^2 being the proportion of variance explained coefficient from the linear regression. It is also possible to use nonlinear adjustment. i.e., instead of allowing only linear adjustment as in (3), we can minimize variance in a more general functional space. Define

$$\widehat{Y}_{cv} = \overline{Y} - \overline{f(X)} + \mathbb{E}(f(X)), \tag{6}$$

and then try to minimize the variance of (6). It can be shown that the regression function $\mathbb{E}(Y|X)$ gives the optimal $f(X)$.

3.2.2 Control Variates in Online Experimentation

Utilizing control variates to reduce variance is a very common technique. The difficulty of applying it boils down to finding a control variate X that is highly correlated with Y and at the same time has known $\mathbb{E}(X)$.

Although in general it is not easy to find control variate X with known $\mathbb{E}(X^{(t)})$ and $\mathbb{E}(X^{(c)})$, a key observation is that $\mathbb{E}(X^{(t)}) - \mathbb{E}(X^{(c)}) = 0$ in the pre-experiment period because we have not yet introduced any treatment effect. By using only information from before the launch of the experiment to construct the control variate, the randomization between treatment and control ensures that we have $\mathbb{E}X^{(t)} = \mathbb{E}X^{(c)}$.

Given $\mathbb{E}X^{(t)} - \mathbb{E}X^{(c)} = 0$, it is easy to see the delta

$$\Delta_{cv} = \widehat{Y}_{cv}^{(t)} - \widehat{Y}_{cv}^{(c)} \tag{7}$$

is an unbiased estimator of $\delta = \mathbb{E}(\Delta)$. Notice how Δ_{cv} does not depend on the unknown $\mathbb{E}(X^{(t)})$ and $\mathbb{E}(X^{(c)})$ at all as they cancel each other. With the optimal choice of θ from Eq (4), we have that Δ_{cv} reduces variance by a factor of ρ^2 compared to Δ, i.e.

$$\text{var}(\Delta_{cv}) = \text{var}(\Delta)(1 - \rho^2).$$

To achieve a large correlation and hence better variance reduction, an obvious approach is to choose X to be the same as Y, which naturally leads to using the same variable during pre-experiment observation window as the control variate. As we will see in the empirical results in Section 5, this indeed turns out to be the most effective choice we found for control variates.

There is a slight subtlety that's worth pointing out. The pair (Y, X) may have different distributions in treatment and control when there is an experiment effect. For Δ_{cv} to be unbiased, the same θ has to be used for both control and treatment. The simplest way to estimate it is from the pooled population of control and treatment. The impact on variance reduction will likely be negligible. In the general nonlinear control covariates case, we should use the same functional form in both $\widehat{Y}_{cv}^{(t)}$ and $\widehat{Y}_{cv}^{(c)}$.

APPENDIX

A. CONTROL VARIATES AS AN EXTENSION OF STRATIFICATION

Here we show that when the covariates are categorical, stratification and control variates produce identical results.

For clarity and simplicity, we assume X is binary with values 1 and 0. Let $w = \mathbb{E}(X)$. The two estimates are

$$\begin{aligned} \widehat{Y}_{strat} &= w\overline{Y}_1 + (1-w)\overline{Y}_0, \\ \widehat{Y}_{cv} &= \overline{Y} - \hat{\theta}\overline{X} + \hat{\theta}w, \end{aligned}$$

where $\overline{Y}_1$ denotes the average of Y in the $\{X = 1\}$ stratum and $\hat{\theta} = \widehat{\text{cov}}(Y, X)/\widehat{\text{var}}(X) = \overline{Y}_1 - \overline{Y}_0$. Plugging in the expression for $\hat{\theta}$, we have

$$\begin{aligned} \widehat{Y}_{cv} &= \overline{Y} - (\overline{Y}_1 - \overline{Y}_0)\overline{X} + (\overline{Y}_1 - \overline{Y}_0)w \\ &= (1-\overline{X})\overline{Y}_0 + \overline{Y}_0\overline{X} + (\overline{Y}_1 - \overline{Y}_0)w \\ &= w\overline{Y}_1 + (1-w)\overline{Y}_0 = \widehat{Y}_{strat}, \end{aligned}$$

where the second equality follows from the fact that $\overline{Y} = \overline{XY}_1 + (1-\overline{X})\overline{Y}_0$.

To prove for the case with $K > 2$, we construct $K-1$ indicator variables as control variates. With the observation that the coefficients $\hat{\theta}_k = \overline{Y}_k - \overline{Y}_0$, the proof follows the same steps as the binary case outlined above.