Lecture 6: Variance reduction

Lecturer: Ksenia Kasianova
xeniakasianova@gmail.com

December 11, 2023

Plan

Plan

1) Variance reduction

2) Paired t-test

3) CUPED

4) Stratification

Plan

Main factors when estimating any measure:

1. sample size is finite
2. the estimate should be as accurate as possible.

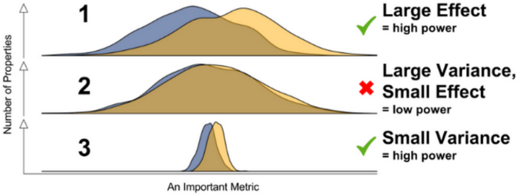

Figure:

Variance reduction: 2 => 3

Variance reduction in General

In mathematics, more specifically in the theory of **Monte Carlo methods**, variance reduction is a procedure used to increase the precision of the estimates obtained for a given simulation, i.e. reducing the variance which limits the precision of the simulation results.

The main variance reduction methods are

- antithetic variates
- control variates
- importance sampling
- stratified sampling
- conditional Monte Carlo
- etc.

General idea: taking advantage of the covariance between the measures in an attempt to reduce the variance of the measure of interest.

Antithetic variates uses measures whose covariance is negatively related, *control variates* uses variates whose covariance is positively related, and *control covariates* utilizes the difference between the other measure and its mean for each measurement.

Variance reduction in Causal inference

1) Simplest example – Paired t-test

For sample X (before treatment) and Y (after treatment),

to test the null hypothesis that the true mean difference is zero, the procedure is as follows:

1. Calculate the difference ($d_i = y_i - x_i$) between the two observations on each pair, making sure you distinguish between positive and negative differences.

2. Calculate the mean difference, $\bar{d}$.

3. Calculate the standard deviation of the differences, s_d, and use this to calculate the standard error of the mean difference, $SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$

4. Calculate the t-statistic, which is given by $T = \frac{\bar{d}}{SE(d)}$. Under the null hypothesis, this statistic follows a t-distribution with $n-1$ degrees of freedom.

5. Use tables of the t-distribution to compare your value for T to the t_{n-1} distribution. This will give the p-value for the paired t-test.

Paired t-test

1) Simplest example – Paired t-test

No reduction: compare the average of the individuals outcomes before and after the treatment.

Paired t-test variance reduction:

If we want the causal effect of the treatment, then we can reduce variance by controlling for the individual.

We do this by measuring each individual's pre-treatment measure, i.e. controlling for each individual's pre-treatment measure.

The covariance between each individual's pre-treatment and post-treatment measure is utilized to reduce the variance of our estimation of the treatment effect.

CUPED (Controlled-experiment Using Pre-Experiment Data)

2) CUPED (Controlled-experiment Using Pre-Experiment Data)

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.

– Probably the most used variance reduction technique in A/B testing in the tech industry.

– Special case of applying the technique of **control variates** to the A/B testing set-up.

CUPED (Controlled-experiment Using Pre-Experiment Data)

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, $\Delta = \mathbb{E}\left[Y_i(T) - Y_i(C)\right]$.

The most commonly used estimator for this is the **difference-in-means estimator**

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

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.

CUPED (Controlled-experiment Using Pre-Experiment Data)

Key idea

Imagine that on top of collecting metric values $Y_1, Y_2, \ldots, Y_{nt}$ in the treatment group, we also collected pre-experiment values on another (real-valued) variable $X_1, X_2, \ldots, X_{nt}$.

Let's also assume that we know the mean of X (which denote by $\mathbb{E}[X]$).

For any fixed parameter θ, we have

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

Hence,

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

is an unbiased estimator for $\mathbb{E}\left[Y_i(T)\right]$.

Additionally, the variance of $\widetilde{Y}_T$ is minimized when $\theta = \operatorname{Cov}(Y, X)/\operatorname{Var}(X)$, and at this value of θ, we have

$$\operatorname{Var}\left(\tilde{Y}_T\right) = \left(1-\rho^2\right)\operatorname{Var}\left(\bar{Y}_T\right) \leq \operatorname{Var}\left(\bar{Y}_T\right)$$

where ρ is the correlation between Y and X.

CUPED (Controlled-experiment Using Pre-Experiment Data)

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

1) We don't know the value of $\theta = \operatorname{Cov}(Y, X)/\operatorname{Var}(X)$.

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 $\mathbb{E}\left[\hat{\theta}\bar{X}_T\right] \neq \theta \mathbb{E}[X]$ in general: both $\hat{\theta}$ and $\bar{X}_T$ 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.)

CUPED (Controlled-experiment Using Pre-Experiment Data)

2) 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 $\bar{Y}_T$. 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 $\mathbb{E}\left[Y_i(T)\right]$ but $\Delta = \mathbb{E}\left[Y_i(T)\right] - \mathbb{E}\left[Y_i(C)\right]$. Hence,

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

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

CUPED (Controlled-experiment Using Pre-Experiment Data)

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

$$\text{Var}\left(\widetilde{Y}_T\right) = \left(1-\rho^2\right)\text{Var}\left(\bar{Y}_T\right) \leq \text{Var}\left(\bar{Y}_T\right)$$

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.

Caution! X must not be affected by the experiment's treatment.

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.

CUPED (Controlled-experiment Using Pre-Experiment Data)

1. Take a random variable X independent of Y
2. Let's imagine the new metric as the difference between Y and θX.

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

3. Variance is calculated using the formula:

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

4. Variance is minimized when:

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

Final variance

$$\operatorname{var}_{srs}\left(Y_{CUPED}\right)_{\min} = \operatorname{var}_{srs}(Y)\left(1 - \rho^2\right)$$

CUPED (Controlled-experiment Using Pre-Experiment Data)

Recommendations [Deng, et. al., 2013]

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.

Stratification

3) Stratification

Idea: (1) divide the sampling region into strata, (2) sample within each stratum separately and (3) combine results from individual strata together to give an overall estimate.

Mathematically, we want to estimate $\mathbb{E}(Y)$, where Y is the variable of interest.

Assume we can divide the sampling region of Y into K subregions (strata) with w_k the probability that Y falls into the k th stratum, $k = 1, \ldots, K$.

If we fix the number of points sampled from the k th stratum to be $n_k = n \cdot w_k$, we can define a stratified average to be

$$\widehat{Y}_{\text{strat}} = \sum_{k=1}^{K} p_k \bar{Y}_k,$$

where $\bar{Y}_k$ is the average within the k th stratum.

The stratified average $\widehat{Y}_{\text{strat}}$ and the standard average $\bar{Y}$ have the same expected value but the former gives a smaller variance when the means are different across the strata.

Stratification

The **intuition** is that the variance of $\bar{Y}$ can be decomposed into the within-strata variance and the between-strata variance, and the latter is removed through stratification.

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,

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

where $\left(\mu_k, \sigma_k^2\right)$ denote the mean and variance for users in the k th stratum.

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.

Stratification

Consider two estimates of the population mean. The first is the standard simple sample average denoted as $\bar{Y}$. It is defined as

$$\bar{Y} = \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n_k} Y_{kj}.$$

The second is a weighted average denoted as $\hat{Y}_{\text{strat.}}$. It is defined as

$$\hat{Y}_{\text{strat}} = \sum_{k=1}^{K} p_k \bar{Y}_k,$$

where p_k is defined above and $\bar{Y}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} Y_{kj}$ is the average of the business metric for users from the k th stratum.

Under stratified sampling the estimates are the same,

$$\begin{aligned}\sum_{k=1}^{K} p_k \bar{Y}_k &= \sum_{k=1}^{K} p_k \frac{1}{n_k} \sum_{j=1}^{n_k} Y_{kj} \\ &= \sum_{k=1}^{K} \frac{n_k}{n} \frac{1}{n_k} \sum_{j=1}^{n_k} Y_{kj} \\ &= \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n_k} Y_{kj}.\end{aligned}$$

Stratification

Stratified sampling:

1) The estimate is unbiased under stratified sampling.

$$E_{\text{strat}}\left(\hat{Y}_{\text{strat}}\right)=\sum_{k=1}^{K} p_k E_{\text{strat}}\left(\bar{Y}_k\right)=\sum_{k=1}^{K} p_k \mu_k=\mu .$$

2) The variance of the estimate under stratified sampling is

$$\begin{aligned}\operatorname{var}_{\text{strat}}\left(\hat{Y}_{\text{strat}}\right) &=\sum_{k=1}^{K} p_k^2 \operatorname{var}_{\text{strat}}\left(\bar{Y}_k\right) \\ &=\sum_{k=1}^{K} \frac{n_k^2}{n^2} \frac{1}{n_k} \sigma_k^2 \\ &=\frac{1}{n} \sum_{k=1}^{K} p_k \sigma_k^2 .\end{aligned}$$

The first equation holds because sampling from the K strata is done independently from each other.

Stratification

Simple random sampling

1) The estimate is unbiased shown as follows

$$\begin{aligned} E_{srs}(\bar{Y}) &= E_{srs}\left(\frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{n_k} Y_{kj}\right) \\ &= \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{n_k} E_{srs}\left(Y_{kj}\right) \\ &= \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{n_k} \mu = \frac{1}{n} n\mu = \mu. \end{aligned}$$

2) ·The variance under simple random sampling is derived as follows.

$$\begin{aligned} \text{var}_{srs}(\bar{Y}) &= \text{var}_{srs}\left(\frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{n_k} Y_{kj}\right) \\ &= \frac{1}{n^2}\sum_{k=1}^{K}\sum_{j=1}^{n_k} \text{var}_{srs}\left(Y_{kj}\right) = \frac{1}{n^2} n\sigma^2 = \frac{1}{n}\sigma^2 \end{aligned}$$

Stratification

Note that $\mathrm{var}_{srs}\left(Y_{kj}\right) = \sigma^2$ because Y_{kj} are all random samples under simple random sampling from the distribution of Y.

Let Z denote the stratum number of a random observation from the distribution of Y under simple random sampling. Note that Z is a multinomial random variable that takes values $1, \ldots, K$ and $P(Z = k) = p_k$. Then we have

$$
\begin{aligned}
\mathrm{var}_{srs}(Y) &= E_{srs}\left(\mathrm{var}_{srs}(Y \mid Z)\right) + \mathrm{var}_{srs}\left(E_{srs}(Y \mid Z)\right) \\
&= E_{srs}\left(\sum_{k=1}^{K} \sigma_k^2 I(Z = k)\right) + \mathrm{var}_{srs}\left(\sum_{k=1}^{K} \mu_k I(Z = k)\right) \\
&= \sum_{k=1}^{K} \sigma_k^2 E_{srs}(I(Z = k)) + E_{srs}\left(\sum_{k=1}^{K} \mu_k I(Z = k)\right)^2 \\
&- \left(E_{srs}\left(\sum_{k=1}^{K} \mu_k I(Z = k)\right)\right)^2 \\
&= \sum_{k=1}^{K} \sigma_k^2 p_k + \sum_{k=1}^{K} \mu_k^2 p_k - \mu^2 \\
&= \sum_{k=1}^{K} \sigma_k^2 p_k + \sum_{k=1}^{K} p_k \left(\mu_k - \mu\right)^2,
\end{aligned}
$$

where $I(Z = k)$ is an indicator variable with value 1 if $Z = k$ and 0 otherwise. Combing all, we have

$$
\mathrm{var}_{srs}(\bar{Y}) = \frac{1}{n} \sum_{k=1}^{K} p_k \sigma_k^2 + \frac{1}{n} \sum_{k=1}^{K} p_k \left(\mu_k - \mu\right)^2
$$

Stratification

Summary of comparison

– Both estimates are unbiased.

– The variance of the estimate in stratified sampling is smaller than that in simple random sampling

$$\begin{aligned}\operatorname{var}(\bar{Y}_{srs}) &= \sum_{k=1}^{K} \frac{p_k}{n} \sigma_k^2 + \sum_{k=1}^{K} \frac{p_k}{n} (\mu_k - \mu)^2 \\ &\geq \sum_{k=1}^{K} \frac{p_k}{n} \sigma_k^2 = \operatorname{var}\left(\widehat{Y}_{strat}\right)\end{aligned}$$

The intuition:

Variance of SRS estimate can be decomposed into within-strata variance and between-strata variance.

Stratified sampling achieves variance reduction by removing the between-strata variance.

Post-stratification

Post stratification assumes simple random sampling but uses the estimate in

$$\hat{Y}_{\text{strat}} = \sum_{k=1}^{K} p_k \bar{Y}_k$$

$$\bar{Y} = \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n_k} Y_{kj}.$$

Note that, when simple random sampling is used, these estimates are different.

This is because the sample size n_k from the k th stratum is not necessarily equal to np_k under simple random sampling.

In fact, $n_1, \ldots, n_K$ are all random under simple random sampling.

The intuition behind post stratification is very simple.

The weighted average gives more weights to observations from the strata that are under-represented in the sample.

Thus if a sample is *badly balanced* for some covariate such as signup country, the weighted average estimate automatically corrects for it.

Post-stratification

We now sketch the derivation of the variance of the estimate under simple random sampling.

$$\begin{aligned}
\mathrm{var}_{srs}(\bar{Y}_{strat}) &= E_{\mathrm{srs}}\left(\hat{\mathrm{var}}_{\mathrm{srs}}\left(\hat{Y}_{\mathrm{strat}} \mid n_1, \ldots, n_K\right)\right) \\
&+ \mathrm{var}_{srs}\left(E_{strat} \mid n_1, \ldots n_K\right)) \\
&= E_{srs}\left(\sum_{k=1}^{K} p_k^2 \, \mathrm{var}_{srs}\left(\bar{Y}_k \mid n_k\right)\right) + \mathrm{var}_{srs}\left(\sum_{k=1}^{K} p_k \mu_k\right) \\
&= E_{srs}\left(\sum_{k=1}^{K} p_k^2 \frac{1}{n_k} \sigma_k^2\right) + \mathrm{var}_{srs}(\mu) \\
&= \sum_{k=1}^{K} p_k^2 \sigma_k^2 E_{srs}\left(\frac{1}{n_k}\right)
\end{aligned}$$

What is to calculate $E_{srs}\left(\frac{1}{n_k}\right)$, where $k = 1, \ldots, K$.

Post-stratification

The proof is technical and based on Taylor expansion of $\frac{1}{n_k}$ at $\frac{1}{np_k}$ and the fact that n_k is Bernoullian variable with mean np_k and variance $np_k(1- p_k)$.

$$\begin{aligned}
E_{srs}\left(\frac{1}{n_k}\right) &= E_{srs}\left(\frac{1}{np_k} + \left(-\frac{1}{n^2p_k^2}\right)(n_k - np_k)\right. \\
&\left. + \frac{1}{n^3p_k^3}(n_k - np_k)^2\right) + o\left(\frac{1}{n^2}\right) \\
&= \frac{1}{np_k} + \frac{1}{n^3p_k^3}E_{srs}(n_k - np_k)^2 + o\left(\frac{1}{n^2}\right) \\
&= \frac{1}{np_k} + \frac{1}{n^3p_k^3}np_k(1 - p_k) + o\left(\frac{1}{n^2}\right) \\
&= \frac{1}{np_k} + \frac{1}{n^2p_k^2}(1 - p_k) + o\left(\frac{1}{n^2}\right),
\end{aligned}$$

Thus, we have

$$\text{var}_{srs}\left(\hat{Y}_{strat}\right) = \frac{1}{n}\sum_{k=1}^{K} p_k\sigma_k^2 + \frac{1}{n^2}\sum_{k=1}^{K}(1 - p_k)\sigma_k^2 + o\left(\frac{1}{n^2}\right).$$

Hence for large enough n, post stratification leads to variance reduction for large enough sample size:

$$\begin{aligned}
\text{var}_{\text{strat}}\left(\hat{Y}_{\text{strat}}\right) &= \text{var}_{\text{srs}}\left(\hat{Y}_{\text{strat}}\right) + O\left(\tfrac{1}{n^2}\right) = \text{var}_{\text{srs}}\left(\bar{Y}\right) + O\left(\tfrac{1}{n}\right), \\
\text{var}_{\text{strat}}\left(\hat{Y}_{\text{strat}}\right) &\leq \text{var}_{\text{srs}}\left(\hat{Y}_{\text{strat}}\right) \leq \text{var}_{\text{srs}}\left(\bar{Y}\right).
\end{aligned}$$

Stratification in pactice

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 **preexperiment variables** to construct strata after all the data are collected.

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,

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

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

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

enjoys the stratified average variance reduction.

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.

Stratification in pactice

SAMPLE SIZES REQUIRED FOR EACH METHOD

	Sample size to estimate a proportion	Sample size to estimate an average
Simple random sampling	$\frac{Z^2 p(1-p)}{e^2}$	$\frac{Z^2\sigma^2}{e^2}$
Proportional stratified sampling	$\frac{Z^2 \sum_{h=1}^{L} W_h p_h \left(1-p_h\right)}{e^2}$	$\frac{Z^2 \sum_{h=1}^{L} W_h \sigma_h^2}{e^2}$
Best stratified sampling	$\frac{Z^2 \left(\sum_{h=1}^{L} W_h \sqrt{p_h \left(1-p_h\right)}\right)^2}{e^2}$	$\frac{Z^2 \left(\sum_{h=1}^{L} W_h \sigma_h\right)^2}{e^2}$

where Z – quantiles of the Gaussian distribution, L is the number of strata, e is the accepted margin of error; σ^2 is the variance of the data within the total population. σ_h^2 is the variance within every stratum.

- p is the proportion of the total population that we are trying to determine (e.g. the percent of the Mexican population that smokes). p_h represents that proportion within each stratum.

- W_h is the stratum's weight within the sample (the size of the stratum with respect to the whole sample).

Proportional – W_h is equal to the proportion represented by that stratum in the population. *Optimal* – W_h is calculated based on the dispersion within each stratum.

CUPED vs Stratification

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

$$\widehat{Y}_{\text{strat}} = w\bar{Y}_1 + (1-w)\bar{Y}_0$$

$$\widehat{Y}_{cv} = \bar{Y} - \hat{\theta}\bar{X} + \hat{\theta}w$$

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

$$\begin{aligned}\widehat{Y}_{cv} &= \bar{Y} - \left(\bar{Y}_1 - \bar{Y}_0\right)\bar{X} + \left(\bar{Y}_1 - \bar{Y}_0\right)w \\ &= (1-\bar{X})\bar{Y}_0 + \bar{Y}_0\bar{X} + \left(\bar{Y}_1 - \bar{Y}_0\right)w \\ &= w\bar{Y}_1 + (1-w)\bar{Y}_0 = \widehat{Y}_{\text{strat}}\end{aligned}$$

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

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

Count metrics vs ratio metrics

Count metrics

For count metrics, the unit of analysis is the same as the unit of randomization. For example, if the unit of analysis is a user, count metrics would include *revenue per user, clicks per user, etc.*

In mathematical notation, let Y denote variable of interest. The count metric we want to estimate (i.e. the estimand) is $\widehat{M} = \frac{\sum_i Y_i}{n}$

Ratio metrics

For ratio metrics, the unit of analysis is at a more granular level than the unit of randomization. For example, when the unit of randomization is a user, ratio metrics would include *revenue per session, clicks per page view, etc.*

Let Y denote the variable of interest, and let Z denote the number of units of analysis for this randomization unit. The ratio metric we want to estimate is

$$R = \frac{\mathbb{E}[Y]}{\mathbb{E}[Z]}.$$

The most common way to estimate this is to replace both the numerator and denominator with the respective sample means:

$$\widehat{R} = \frac{\sum_i Y_i/n}{\sum_i Z_i/n} = \frac{\sum_i Y_i}{\sum_i Z_i}$$

CUPED and CTR

1. User metrics

$$OEC_A = \text{avg}_{u \in A} X(u)$$

Example:

$$\frac{3+0+6+\ldots+5+6}{N} = \frac{3+0+6+\ldots+5+6}{1+1+1+\ldots+1+1}$$

2. Ratio metrics:

$$OEC_A = \frac{\sum_{u \in A} X(u)}{\sum_{u \in A} Y(u)}$$

Example:

$$\frac{3+0+6+\ldots+5+6}{9+8+7+\ldots+7+8}$$

CUPED and CTR

How to conduct t-test for Ratio metrics? For T-test you need to be able to calculate,

$$R = \frac{X}{Y} \quad \mathrm{E}[R] = ? \quad \mathrm{Var}(R) = ?$$

Linearization Ratio

Taylor series up to 1st order:

$$f(a, b) + (x - a)f_x(a, b) + (y - b)f_y(a, b)$$

Expansion of $R = \frac{X}{Y}$ at point $(\mathrm{E}[X], \mathrm{E}[Y])$:

$$Z = \frac{\mathrm{E}[X]}{\mathrm{E}[Y]} + \frac{1}{\mathrm{E}[Y]}\left(X - \frac{\mathrm{E}[X]}{\mathrm{E}[Y]}Y\right)$$

$$E[Z] \approx \mathrm{E}[R] \quad \mathrm{Var}(Z) \approx \mathrm{Var}(R)$$

This formula converts two samples (numerator and denominator) into one, preserving the mean and variance (asymptotically), which allows the use of classical tests.

CUPED and CTR

Consider $Z = \frac{X}{Y}$ with

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}\right)$$

Consider random variables R and S where S either has no mass at 0 (discrete) or has support $[0, \infty)$. Let $G = g(R, S) = R/S$.

Find approximations for EG and $\mathrm{Var}(G)$ using Taylor expansions of $g()$.

For any $f(x, y)$, the bivariate first order Taylor expansion about any $\boldsymbol{\theta} = (\theta_x, \theta_y)$ is

$$f(x, y) = f(\boldsymbol{\theta}) + f_x'(\boldsymbol{\theta})\,(x - \theta_x) + f_y'(\boldsymbol{\theta})\,(y - \theta_y) + \mathbf{R}$$

where R is a remainder of smaller order than the terms in the equation.

CUPED and CTR

Switching to random variables with finite means $EX \equiv \mu_x$ and $EY \equiv \mu_y$, we can choose the expansion point to be $\theta = (\mu_\mathbf{x}, \mu_\mathbf{y})$. In that case the first order Taylor series approximation for $f(X, Y)$ is

$$f(X, Y) = f(\boldsymbol{\theta}) + f_x'(\boldsymbol{\theta})\,(X - \mu_x) + f_y'(\boldsymbol{\theta})\,(Y - \mu_y) + R$$

The approximation for $E(f(X, Y))$ is therefore

$$\begin{aligned}
E(f(X, Y)) &= E\left[f(\boldsymbol{\theta}) + f_x'(\boldsymbol{\theta})\,(X - \mu_x) + f_y'(\boldsymbol{\theta})\,(Y - \mu_y) + R\right] \\
&\approx E[f(\boldsymbol{\theta})] + E\left[f_x'(\boldsymbol{\theta})\,(X - \mu_x)\right] + E\left[f_y'(\boldsymbol{\theta})\,(Y - \mu_y)\right] \\
&= E[f(\boldsymbol{\theta})] + f_x'(\boldsymbol{\theta})E\left[(X - \mu_x)\right] + f_y'(\boldsymbol{\theta})E\left[(Y - \mu_y)\right] \\
&= E[f(\boldsymbol{\theta})] + 0 + 0 \\
&= f\left(\mu_x, \mu_y\right)
\end{aligned}$$

For our example where $f(x, y) = x/y$ the approximation is $E(X/Y) = E(f(X, Y)) = f\left(\mu_x, \mu_y\right) = \mu_x/\mu_y$

CUPED and CTR

Click-through rate (CTR) is the ratio of clicks on a specific link to the number of times a page, email, or advertisement is shown.

To achieve variance reduction for non-user level metrics, we need to incorporate **delta method**.

We use CTR as an example and derive for the control variates formulation since it's more general.

Let n be the number of users (non-random). Denote $Y_{i,j}$ the number of clicks on user i 's j th page-view during the experiment and $X_{i,k}$ the number of clicks on user i 's k th page-view during the pre-experiment period. Let N_i and M_i be the numbers of page-views from user i during the experiment and pre-experiment respectively. The estimate for CTR using $X_{i,j}$ as the control variate becomes

$$\begin{aligned}\widehat{Y}_{cv} &= \frac{\sum_{i,j} Y_{i,j}}{\sum_{i,j} 1} - \theta \frac{\sum_{i,k} X_{i,k}}{\sum_{i,k} 1} + \theta \mathbb{E}\left(X_{i,k}\right) \\ &= \frac{\sum_i Y_{i,+}}{\sum_i N_i} - \theta \frac{\sum_i X_{i,+}}{\sum_i M_i} + \theta \mathbb{E}\left(X_{i,j}\right),\end{aligned}$$

where $Y_{i,+} = \sum_j Y_{i,j}$ is the total number of clicks from user i. Similar notation applies to $X_{i,+}$.

CUPED and CTR

Following the same derivation as for count metrics, we know var $\left(\widehat{Y}_{cv}\right)$ is minimized at

$$
\begin{aligned}
\theta &= \operatorname{cov}\left(\frac{\sum_i Y_{i,+}}{\sum_i N_i}, \frac{\sum_i X_{i,+}}{\sum_i M_i}\right) / \operatorname{var}\left(\frac{\sum_i X_{i,+}}{\sum_i M_i}\right) \\
&\doteq \operatorname{cov}\left(\frac{\bar{Y}}{\mu_N} - \frac{\mu_Y \bar{N}}{\mu_N^2} - \frac{\mu_Y}{\mu_N}, \frac{\bar{X}}{\mu_M} - \frac{\mu_X \bar{M}}{\mu_M^2} - \frac{\mu_X}{\mu_M}\right) / \operatorname{var}\left(\frac{\bar{X}}{\mu_M} - \frac{\mu_X \bar{M}}{\mu_M^2} - \frac{\mu_X}{\mu_M}\right) \\
&= \operatorname{cov}\left(\frac{\bar{Y}}{\mu_N} - \frac{\mu_Y \bar{N}}{\mu_N^2}, \frac{\bar{X}}{\mu_M} - \frac{\mu_X \bar{M}}{\mu_M^2}\right) / \operatorname{var}\left(\frac{\bar{X}}{\mu_M} - \frac{\mu_X \bar{M}}{\mu_M^2}\right)
\end{aligned}
$$

where the second equality follows from using Taylor expansion to linearize the ratios and $\bar{Y} = \frac{1}{n}\sum_i Y_{i,+}$ with $\mu_Y = \mathbb{E}(\bar{Y})$ (similarly for μ_X, μ_N and μ_M).

Because the user is the randomization unit and user level observations are i.i.d., we have

$$\sqrt{n}(\bar{Y}, \bar{N}, \bar{X}, \bar{M}) \Rightarrow N(\mu, \Sigma),$$

following a multivariate normal distribution with mean vector μ and covariance matrix Σ easily estimated from the i.i.d. samples. It is now straight forward to estimate θ using

$$\theta = \left(\beta_1^T \Sigma \beta_2\right) / \left(\beta_2^T \Sigma \beta_2\right),$$

where $\beta_1 = \left(1/\mu_N, -\mu_Y/\mu_N^2, 0, 0\right)^T$ and $\beta_2 = \left(0, 0, 1/\mu_M, -\mu_X/\mu_M^2\right)^T$.

We can easily see that the derivation works generally for various combinations. The metric can be at user level while the covariate can be at page-view level, etc.

CUPED and Frisch-Waugh-Lovell THM

Frisch-Waugh-Lovell THM:

The theorem states that when estimating a linear model of the form

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

then, the following estimators of β_1 are equivalent:

- the OLS estimator obtained by regressing y on x_1 and x_2

- the OLS estimator obtained by regressing y on $\tilde{x}_1$, where $\tilde{x}_1$ is the residual from the regression of x_1 on x_2

- the OLS estimator obtained by regressing $\tilde{y}$ on $\tilde{x}_1$, where $\tilde{y}$ is the residual from the regression of y on x_2

Q: Why Frisch-Waugh?

A: 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:

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

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

CUPED and Frisch-Waugh-Lovell THM

A: 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.

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 assigmments are uncorrelated, i.e. properly randomized, the estimates of the treatment effects will remain unbiased, but the standard errors will be smaller than they were with separate regressions.

CUPED and Frisch-Waugh-Lovell THM

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.

So if you're cunning 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.

Finally, you can also flip on White standar derrors in order to account for treatments whose within-group variance differs from the control, the same as Welch's t-test does.

Summary

Summary:

1) Many other variance reduction methods are developed and productionalized in the tech industry to improve the sensitivity/power of experiments.

– Stratification and post-stratification

– CUPED (controlled-experiment using pre-experiment data)

– Variance-Weighted Estimators

ML-based methods:

– CUPAC (control using predictions as covariates)

– MLRATE (machine learning regression-adjusted treatment effect estimator)

2) CUPED: Using a pre-experiment period of 1-2 weeks works well for variance reduction. Never use covariates that could be affected by the treatment, as this could bias the results.

3) When the covariates are categorical, stratification and control variates produce identical results.

4) Ratio metrics vs count metrics

5) CUPED vs Linear regression