CUPED (Controlled-experiment Using Pre-Experiment Data)

March 15, 2019

Focus on the case of the two-sample t-test. Suppose interested in some metric Y , e.g. Queries per user. To apply the t-test, we assume the observed values of the metric for users in the treatment and control are independent realizations of random variables $Y^{(t)}$ and $Y^{(c)}$. How we think about this - there are two buckets/urns/deck-of-cards and you're pulling samples out from both. Then you apply a t-test based on the test statistic

$$\frac{\overline{Y}^{(t)} - \overline{Y}^{c)}}{\sqrt{var(\overline{Y}^{(t)} - \overline{Y}^{(c)})}} \tag{1}$$

where $\Delta = \overline{Y}^{(t)} - \overline{Y}^{(c)}$ is an unbiased estimator for the shift of the mean and the t-statistic is a normalized version of that estimator. Typically, n will be large here so the normality assumption is not required by the CLT.

We assume the samples are independent, so we have

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

In this framework, the key to variance reduction in the difference is reducing the variance of the means themselves. To do so, compute a corrected estimate of the delta Δ^* that

- is still an unbiasd estimator for the shift in the means, $E(\Delta^*) = \Delta$
- has smaller variance, $var(\Delta^*) < var(\Delta)$

Because of the smaller variance, the test statistic (1) would be larger for the same expected effect size. Hence - improved sensitivity.

Sensitivity of a test

Take a slight detour. Sensitivity here is the same as power - how good the experiment is at detecting an effect, or equivalently the probability of rejecting H_0 when the alternative is true.

Suppose we consider a one-sided test case where we're guessing $H_A : \mu_D = \theta > 0$. Then we have that the power is:

$$\begin{aligned}
B(\theta) &= Pr(T_n > Z_{0.95} | \mu_D = \theta) \\
&\approx Pr(\frac{\overline{D}_n - 0}{\sigma/\sqrt{n}} > 1.64 | \mu_D = \theta) \\
&= Pr(\frac{\overline{D}_n - \theta}{\sigma/\sqrt{n}} > 1.64 - \frac{\theta}{\sigma/\sqrt{n}} | \mu_D = \theta) \\
&= 1 - Pr(\frac{\overline{D}_n - \theta}{\sigma/\sqrt{n}} < 1.64 - \frac{\theta}{\sigma/\sqrt{n}} | \mu_D = \theta) \\
&= 1 - \Phi(1.64 - \frac{\theta}{\sigma/\sqrt{n}} | \mu_D = \theta)
\end{aligned}$$

Put in some numbers to sanity check this.

$n = 0 : \Phi = 0.95, B(\theta) = 0.05$ - if you have no data points at all, the chance of you concluding there's an effect is the same as the significance level of the test. (Think about the test statistic and what it means)

$n \uparrow$: $\Phi \downarrow \implies B(\theta) \uparrow$

The more data points you have, the more powerful your test. But the power only grows with the square root of n!

Variance Reduction

Come from two angles - stratified sampling and control variates.

Stratified sampling

Divide the sampling region into several strata, sample from each stratum and then combine the results from each stratum to get an overall estimate. This estimate usually has a smaller variance than the estimator without stratification.

Mathematically speaking: Originally we want to estimate $E(Y)$. The standard Monte Carlo approach is to simulate $Y_1, Y_2, Y_3, ...Y_n$ and calculate the sample average $\overline{Y}$. Then $\overline{Y}$ is an unbiased estimator of $E(Y)$ and the variance is $Var(\overline{Y}) = \frac{Var(Y)}{n}$.

Thus originally: Use $\overline{Y}$ with variance $\frac{Var(Y)}{n}$.

What about stratified? Assume we can divide the sampling region of Y into K subregions with w_k the probability that Y falls into the kth stratum, $k = 1, 2, ...K$. If we fix the number of points sampled from the kth stratum to be $n_k = n \cdot w_k$. Define the *stratified average*:

$$\hat{Y}_{strat} = \sum_{k=1}^{K} w_k \overline{Y}_k$$

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

$\hat{Y}_{strat}$ is an unbiased estimator, it has the same expected value as $\bar{Y}$:

$$\begin{aligned}
E(\hat{Y}_{strat}) &= E(\sum_{k=1}^{K} w_k \overline{Y}_k) \\
&= \sum_{k=1}^{K} w_k E(\bar{Y}_k) \\
&= \sum_{k=1}^{K} \frac{n_k}{n} E(\bar{Y}_k) \\
&= \sum_{k=1}^{K} \frac{n_k}{n} (\frac{\sum_{i \in k} y_i}{n_k}) \\
&= \sum_{k=1}^{K} \sum_{i \in k} \frac{y_i}{n} \\
&= \sum_{n} \frac{y_i}{n} \\
&= E(\bar{Y})
\end{aligned}$$

But it has a reduced variance. The intuition is, $Var(\bar{Y})$ can be decomposed into within-strata variance and between-strata variance. The latter is removed through stratification. E.g, variance of children's height is large but if you stratify by age, get a much smaller variance within each group.

$$var(\bar{Y}) = \sum_{k=1}^{K} \frac{w_k}{n} \sigma_k^2 + \sum_{k=1}^{K} \frac{w_k}{n} (\mu_k - \mu)^2 \tag{2}$$

$$\geq \sum_{k=1}^{K} \frac{w_k}{n} \sigma_k^2 = var(\hat{Y}_{strat}) \tag{3}$$

We'll prove (3). (2) will unfortunately need to be taken on faith.

$$
\begin{aligned}
var(\hat{Y}_{strat}) &= var(\sum_{k=1}^{K} w_k \bar{Y}_k) \\
&= \sum_{k=1}^{K} var(w_k \bar{Y}_k) \\
&= w_k^2 \sum_{k=1}^{K} var(\bar{Y}_k) \\
&= w_k^2 \sum_{k=1}^{K} var(\frac{\sum_{i \in k} y_i}{n_k}) \\
&= \frac{w_k^2}{n_k} \sum_{k=1}^{K} \sigma_k^2 \\
&= \frac{w_k}{n} \sum_{k=1}^{K} \sigma_k^2
\end{aligned}
$$

□

Example of stratification - If Y_i is the number of queries from a user i , a covariate X_i could be the browser the user used before the experiment started. The estimator $\Delta_{strat} = \hat{Y}_{strat}^{(t)} - \hat{Y}_{strat}^{(c)} = \sum_{k=1}^{K} w_k(\bar{Y}_k^{(t)} - \bar{Y}_k^{(c)})$ enjoys variance reduction.

Control variates

The other angle to reduce variance. Suppose we can simulate another random variable X in addition to Y with known expectation $E(X)$. That is, we have *independent* pairs of $(Y_i, X_i), i = 1, ..., n$. Define

$$\hat{Y}_{cv} = \bar{Y} - \theta \bar{X} + \theta E(X)$$

where θ is any constant. This is an unbiased estimator of $E(Y)$, and it has variance

$$
\begin{aligned}
var(\hat{Y}_{cv}) &= var(\bar{Y} - \theta \bar{X}) \\
&= var(\frac{\sum y}{n} - \frac{\theta}{n} \sum x) \\
&= \frac{1}{n^2} var(\sum y - \theta \sum x) \\
&= \frac{1}{n^2} var(\sum (y - \theta x)) \\
&= \frac{1}{n} var(Y - \theta X) \\
&= \frac{1}{n} \{var(Y) + \theta^2 var(X) - 2\theta cov(Y, X)\}
\end{aligned}
$$

By differentiating with respect to θ, one can find the choice of θ that minimizes this variance:

$$\theta^* = \frac{cov(Y, X)}{var(X)}$$

And with this optimal θ^* , we have

$$
\begin{aligned}
var(\hat{Y}_{cv}) &= \frac{1}{n} \{var(Y) + \theta^2 var(X) - 2\theta cov(Y, X)\} \\
&= \frac{1}{n} \{var(Y) + \frac{cov^2}{var(X)} - 2\frac{cov^2}{var(X)}\} \\
&= \frac{1}{n} \{var(Y) - \frac{cov^2}{var(X)var(Y)} var(Y)\} \\
&= \frac{var(Y)}{n} \{1 - \rho^2\} \\
&= var(\bar{Y})(1 - \rho^2)
\end{aligned}
$$

So the variance is reduced by a factor of ρ^2. The task becomes, find a covariate X that is highly correlated with the metric of interest Y.

In general it isn't easy to find control variate X with known $E(X)^{(t)}$ and known $E(X)^{(c)}$. A key observation is that $E(X)^{(t)} - E(X)^{(c)} = 0$ in the pre-experiment period because there is no treatment effect introduced yet. 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 $EX^{(t)} = EX^{(c)}$. But that means that $\Delta_{cv} = \hat{Y}_{cv}^{(t)} - \hat{Y}_{cv}^{(c)}$ is an unbiased estimator of $\delta = E(\Delta)$ (Slight subtlety here - also means that θ has to be the same for both test and control. The solution is to estimate θ with pooled test and control, i.e. before experiment - but then if you choose the exact same variable, that's as good as saying $\theta = 1$ (?)). Similarly, we see the variance is reduced as well:

$$\begin{aligned} var(\Delta_{cv}) &= var(\hat{Y}_{cv}^{(t)} - \hat{Y}_{cv}^{(c)}) \\ &= var(\hat{Y}_{cv}^{(t)}) + var(\hat{Y}_{t}^{(c)}) \\ &= var(\bar{Y}^{(t)})(1-\rho^2) + var(\bar{Y}^{(c)})(1-\rho^2) \\ &= var(\bar{Y}^{(t)} - \bar{Y}^{(c)})(1-\rho^2) \\ &= var(\Delta)(1-\rho^2) \end{aligned}$$

Thus far we have described two approaches, *control variates* and *stratified sampling*. They both lead to variance reduction, and turns out are actually connected mathematically - one can think of the *control variates* approach as a generalization of *stratified sampling*, that can handle both categorical and numerical covariates.

But what should we use as covariates? Turns out, just use the same variable!