

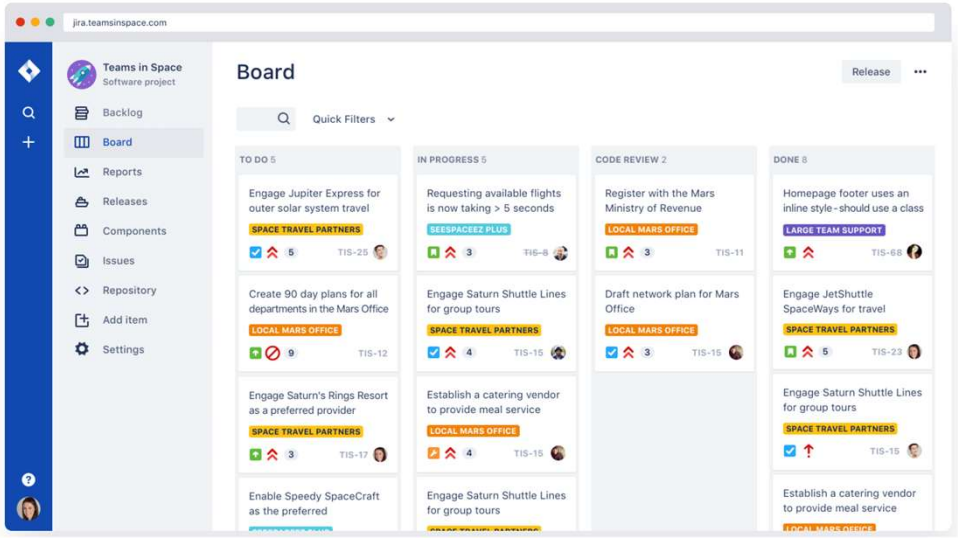
CUPED

A variance reduction technique to increase statistical power

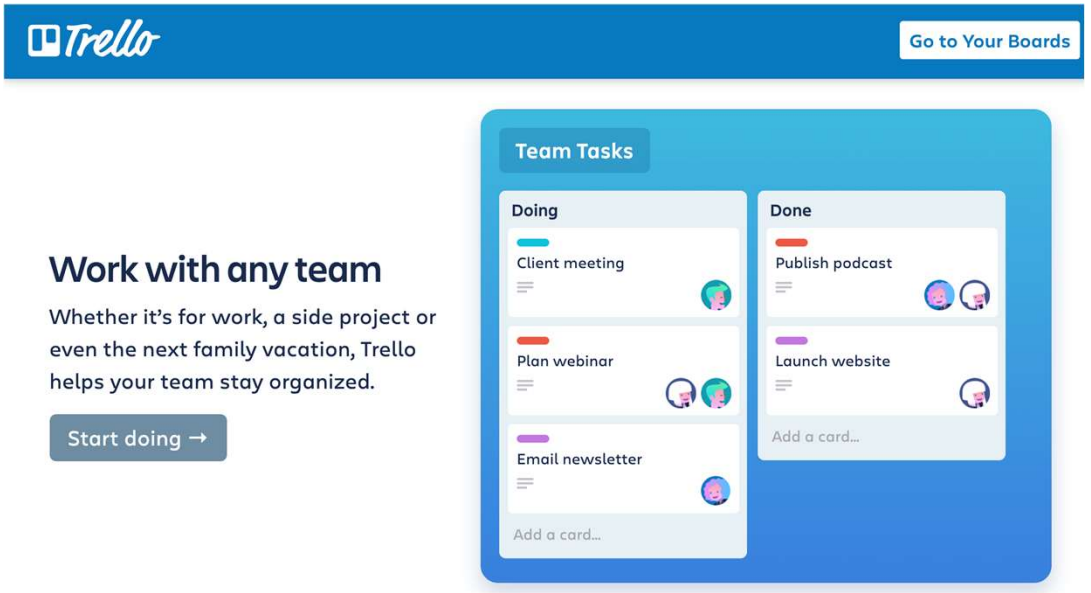


DEREK HO | PRINCIPAL DATA SCIENTIST @ ATLASSIAN

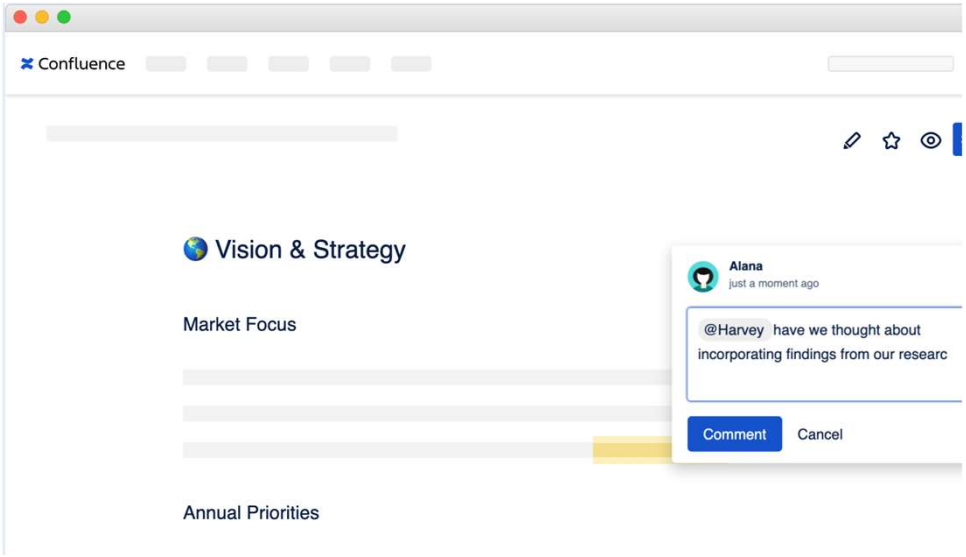
Jira Software



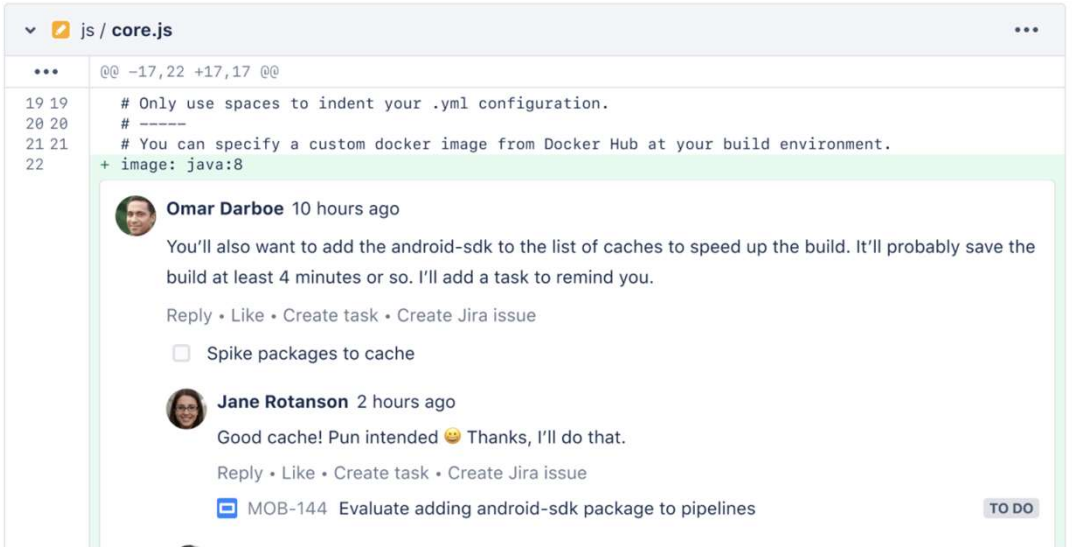
Trello



Confluence



Bitbucket



Talking Point



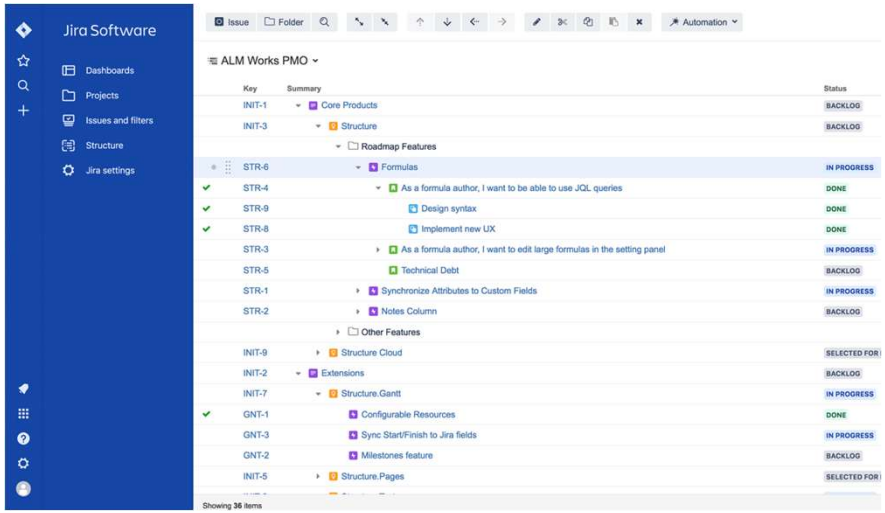
Context



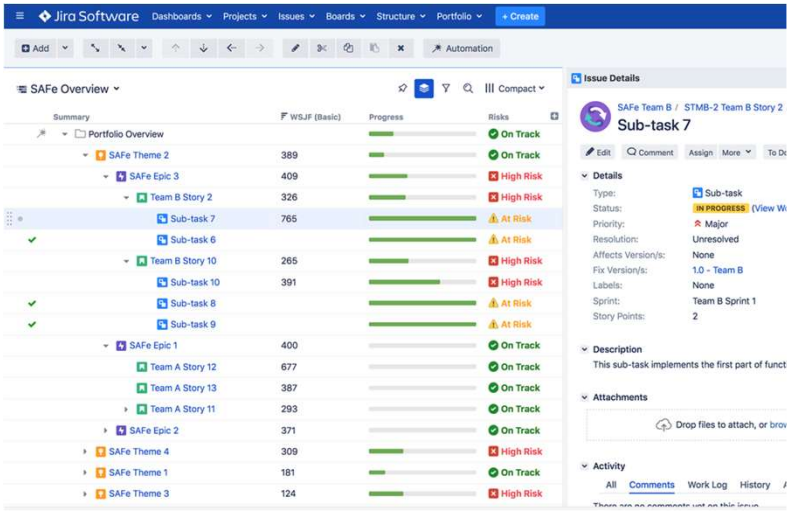
How CUPED works



Factors



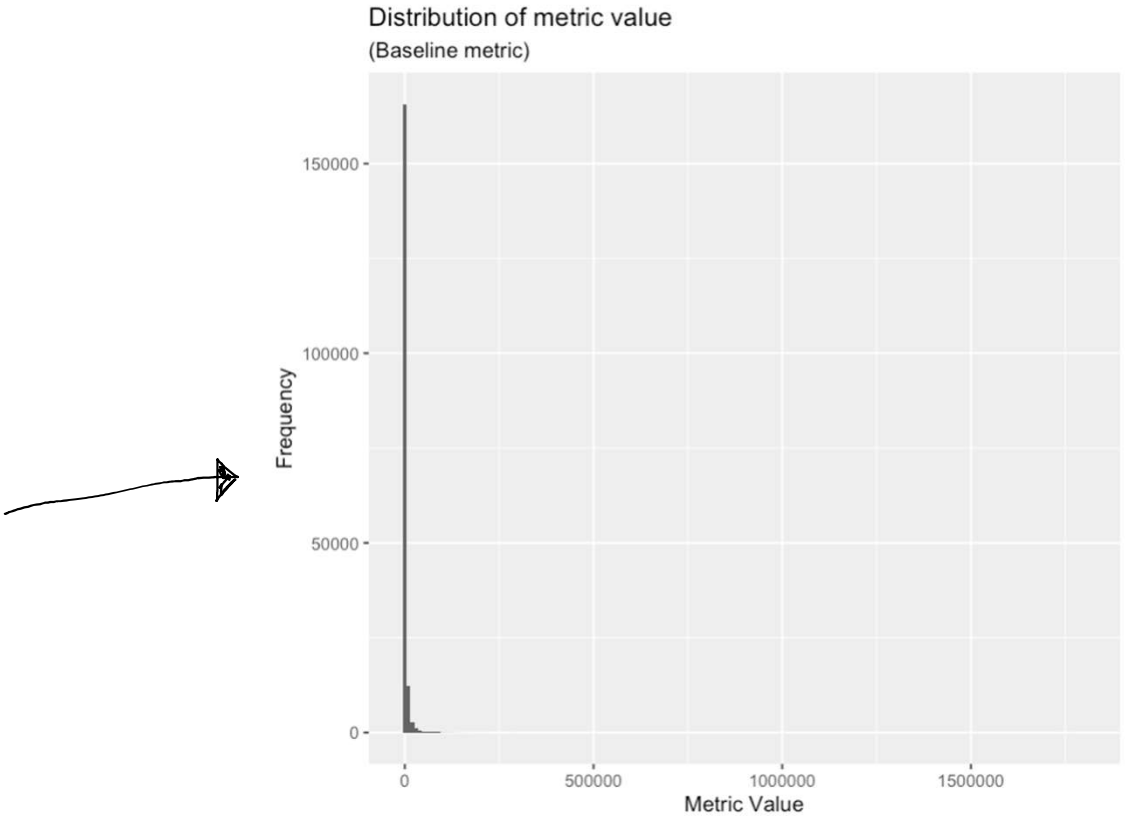
VS



Version **A**
(Control)

Version **B**
(Treatment)

tenantid	pre
001913c1-48dd-41cc-966b-d7de4fd66514	15
0023bb83-c57f-4279-818b-81c909020243	58
008d7113-d150-4354-9239-1ae9c0b585d3	27
00b9804e-70b5-4c71-88a5-665b56d5df9b	7
0112e393-4883-4949-b997-7e7db34c123f	844
012c6876-10b1-4585-bf69-0a685af14a0a	1409
014e8967-a9ad-407c-ae23-8cf41d3d2991	222
0151ffc4-7dc3-456a-90a5-f19212316b6a	4781
016a5d32-92e2-4463-a37c-537f64a15814	20738
019091e8-2953-4b62-9da9-9d3e746330b0	1922
019101fe-6e0e-4bc9-bf9b-7e17c14ab73b	568
01adc561-9cad-4ac5-9a64-a6d1ae7310e8	104
01ccd2a1-ae1a-4a8a-b1fd-7edc04cf57aa	81
01de6db5-c8e7-4f47-868a-0a3aed9c7c8d	57
023f61af-3947-41ed-8a39-db91818f725f	37
02521d91-d433-49b4-87e9-4b4cbfb66921	40
02726b7b-3d54-4203-bf72-68a8663de073	21
0298a6f5-770c-411e-865d-486b7b3eb58b	141
02dd6d1d-fe7c-4ebf-9544-32dc7b4e59b0	392
02f972a4-743d-437e-9dfe-f7c898c47ce2	47
03205ee4-b408-47d5-98c3-f9ad0ee5e3e7	5215
033e1d39-18dd-4f1c-9b46-e53b4d39930f	16
0373824e-cd14-435c-a331-98690800d404	802
0381f98e-c291-4e9b-a519-e2d1d3e80cb2	6950
038de925-a4c4-4a06-aef8-cc265a25ed7c	239
03a4cff7-d3f0-4d30-95d9-5838089a8983	33
03c5801e-0446-49c8-bb48-094cf21c9221	586
03c5d327-8c67-409a-a0b4-7db1a3b1e84c	787
03df99f4-1f3e-43ee-bbd2-8dc8d517f2bd	47



Statistics	Value
Min	1
Median	271
Mean	3,021
Max	1,798,223
Std Dev	33,076

To detect 5% difference in mean, with $\alpha = 0.05$ and $1-\beta = 0.8$ …

752,701

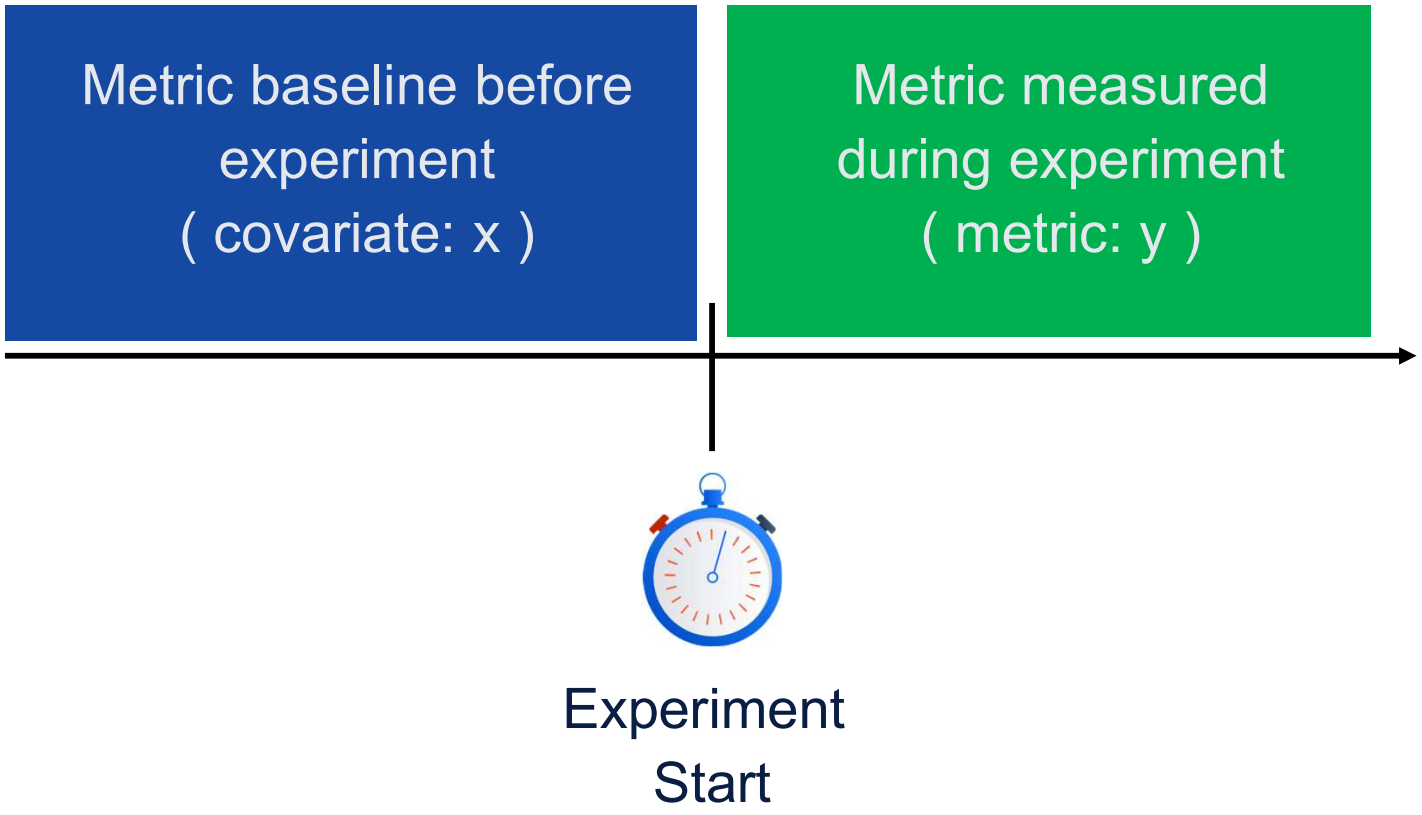
samples per group



Controlled experiment **U**sing **P**re-**E**xperiment **D**ata (CUPED)

How CUPED Works

- Core Idea
- Maths
- Implementation



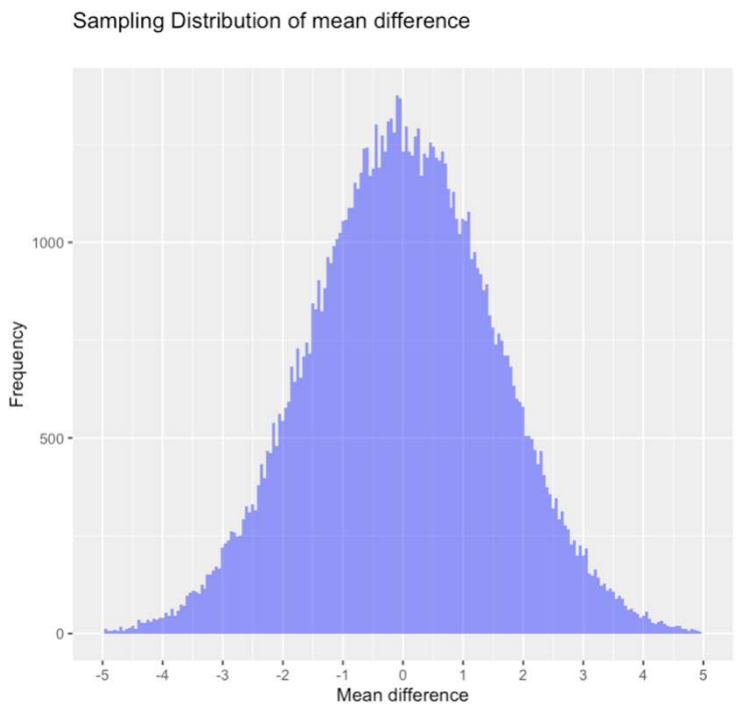
How CUPED Works

Core Idea

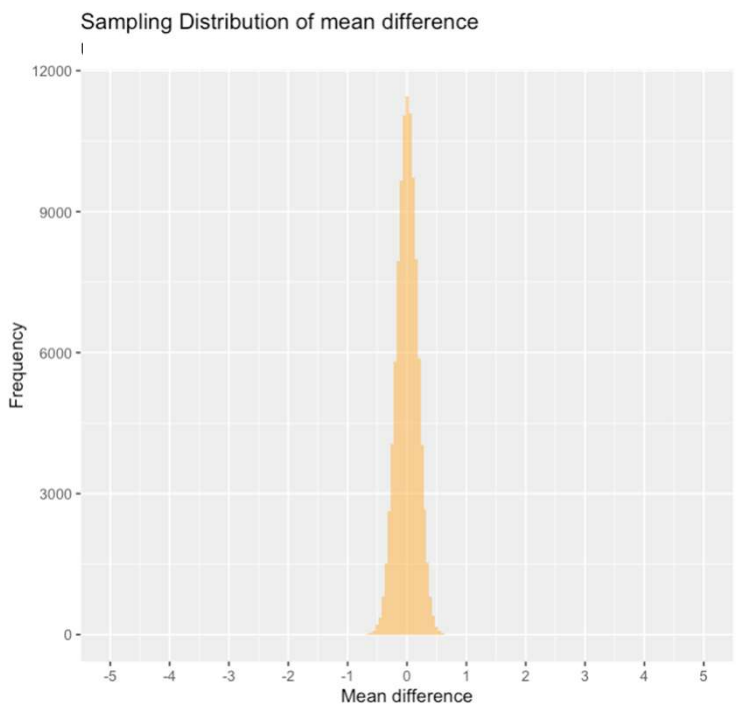
Maths

Implementation

Experiment metric (y)



CUPED-adjusted experiment metric (y*)



Remove variance caused by pre-experiment metric baseline (x)

How CUPED Works

Core Idea

Maths

Implementation

Experiment Metric

$$E(y) = \mu$$

Pre-Experiment Metric baseline

$$E(x) = \tau$$

CUPED-adjusted Experiment Metric

$$y^* = y + c(x - \tau)$$

$$\begin{aligned} E(y^*) &= E(y) + c(E(x) - \tau) \\ &= \mu + c(\tau - \tau) \\ &= \mu \end{aligned}$$

$$c = -\frac{Cov(y,x)}{Var(x)}$$

$$\begin{aligned} Var(y^*) &= Var[y + c(x - \tau)] \\ &= Var(y) + c^2 Var(x) + 2c \cdot Cov(y,x) \\ &= Var(y) + \left(-\frac{Cov(y,x)}{Var(x)}\right)^2 Var(x) + 2\left(-\frac{Cov(y,x)}{Var(x)}\right) \cdot Cov(y,x) \\ &= Var(y) - \frac{Cov^2(y,x)}{Var(x)} \\ &= Var(y) - Var(y)\frac{Cov^2(y,x)}{Var(x) \cdot Var(y)} \\ &= Var(y) \cdot (1 - \rho_{x,y}^2) \\ &< Var(y) \end{aligned}$$

How CUPED Works

Core Idea

Maths

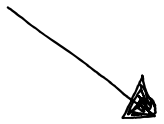
Implementation

tenantid	cohort	pre	post_sim
001913c1-48dd-41cc-966b-d7de4fd66514	ctrl	15	40
0023bb83-c57f-4279-818b-81c909020243	trt	58	107
008d7113-d150-4354-9239-1ae9c0b585d3	ctrl	27	30
00b9804e-70b5-4c71-88a5-665b56d5df9b	ctrl	7	124
0112e393-4883-4949-b997-7e7db34c123f	ctrl	844	3065
012c6876-10b1-4585-bf69-0a685af14a0a	trt	1409	2665
014e8967-a9ad-407c-ae23-8cf41d3d2991	trt	222	53



```
theta = cov(post_sim, pre)/var(pre)

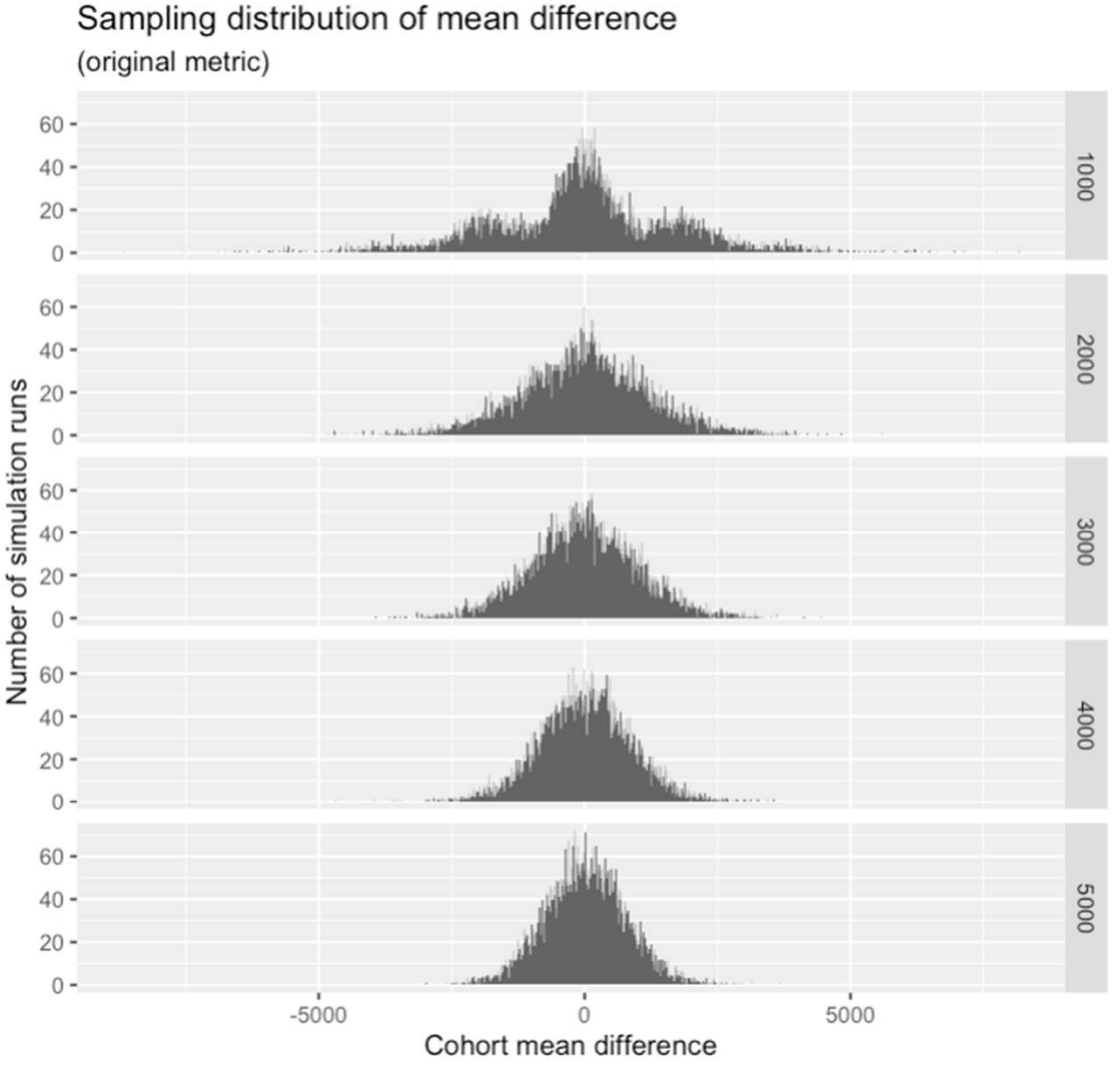
post_sim_cuped = post_sim - theta*(pre - mean(pre))
```

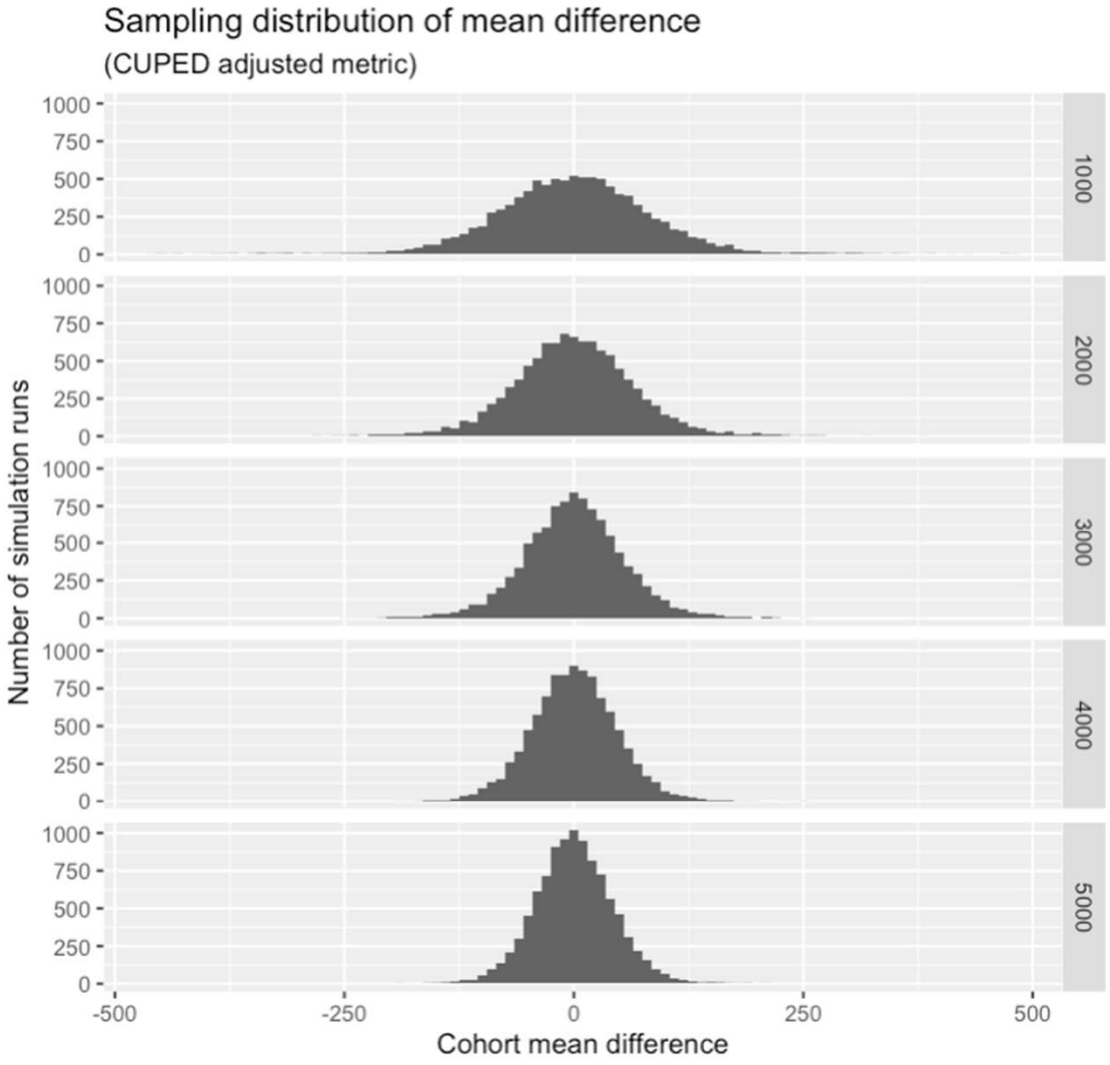


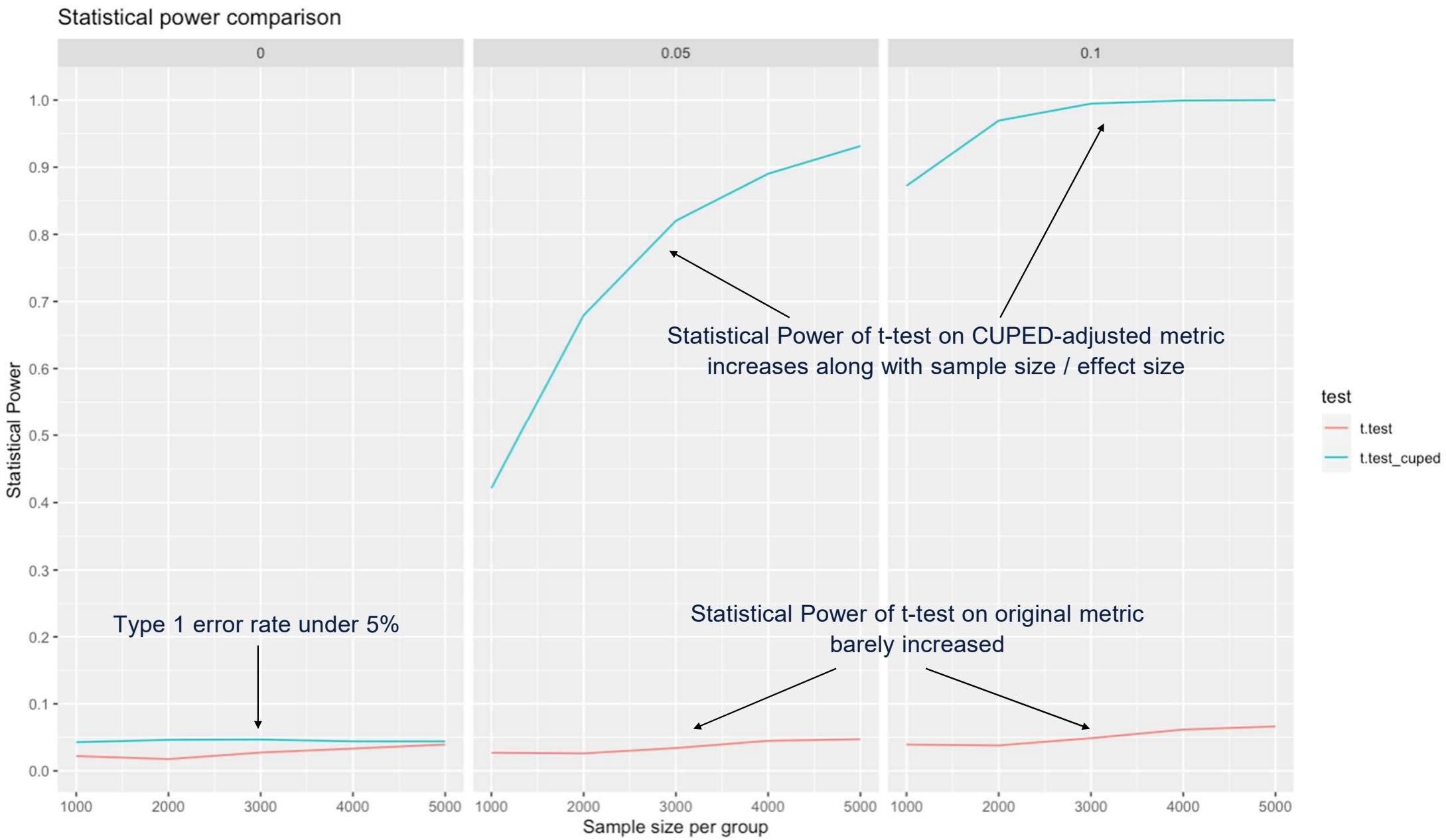
t-test

tenantid	cohort	pre	post_sim	post_sim_cuped
001913c1-48dd-41cc-966b-d7de4fd66514	ctrl	15	40	4467.592466955704
0023bb83-c57f-4279-818b-81c909020243	trt	58	107	4484.059181873294
008d7113-d150-4354-9239-1ae9c0b585d3	ctrl	27	30	4443.490154839683
00b9804e-70b5-4c71-88a5-665b56d5df9b	ctrl	7	124	4560.994008366385
0112e393-4883-4949-b997-7e7db34c123f	ctrl	844	3065	6518.357738273883
012c6876-10b1-4585-bf69-0a685af14a0a	trt	1409	2665	5454.373876144537
014e8967-a9ad-407c-ae23-8cf41d3d2991	trt	222	53	4237.327582954333

Sample Size	Original Metric (post_sim)	CUPED adjusted Metric (post_sim_cuped)
1,000	mean : 3,096 sd : 14,185	mean : 3,096 sd : 1,508
2,000	mean : 3,208 sd : 34,406	mean : 3,208 sd : 2,057
3,000	mean : 2,919 sd : 12,367	mean : 2,919 sd : 1,650
4,000	mean : 3,080 sd : 33,517	mean : 3,080 sd : 2,038
5,000	mean : 3,901 sd : 50,207	mean : 3,901 sd : 1,567


Sampling distribution of mean difference
(original metric)
Number of simulation runs
Cohort mean difference
1000
2000
3000
4000
5000
-5000
0
5000

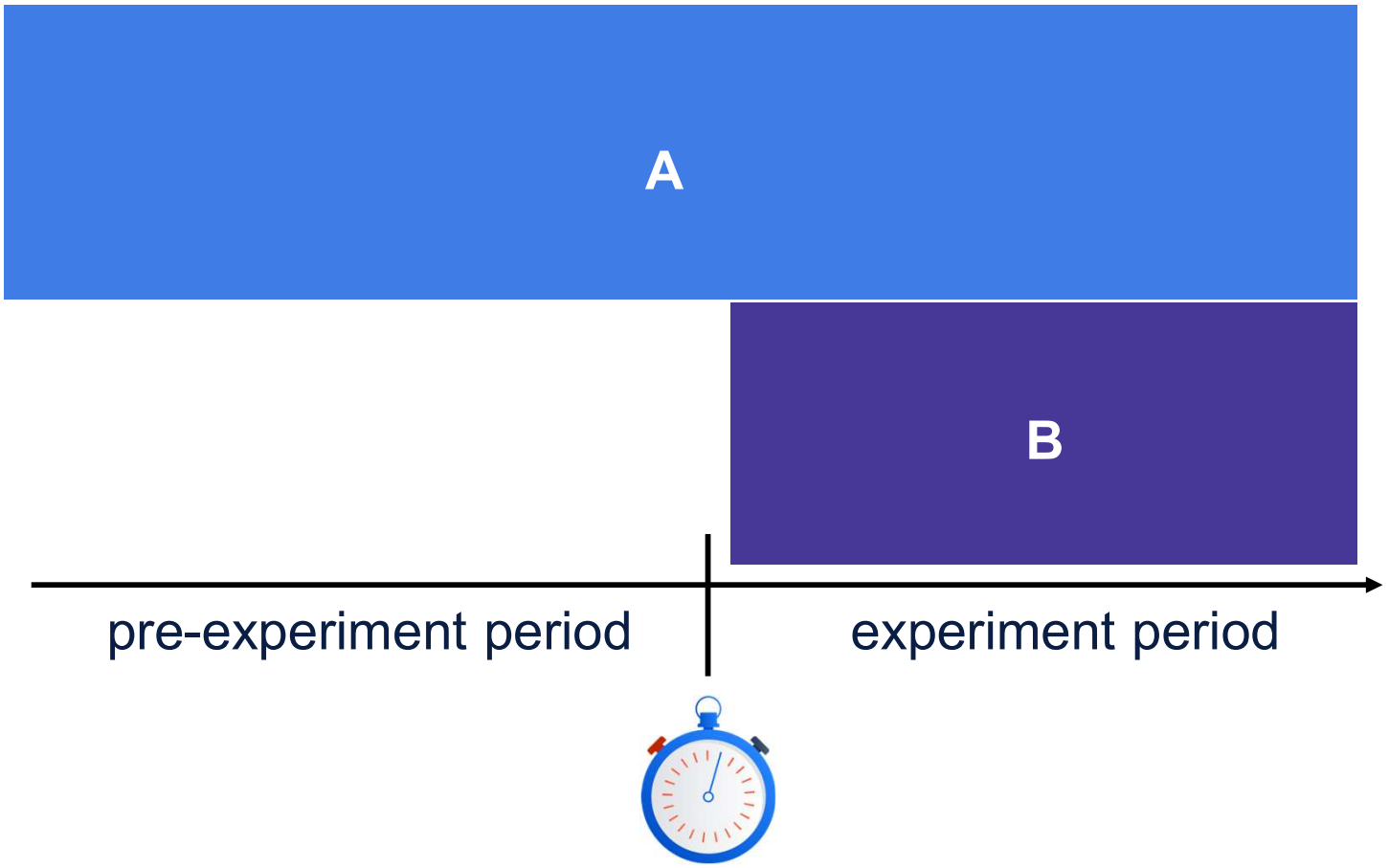

Sampling distribution of mean difference
(CUPED adjusted metric)
Number of simulation runs
Cohort mean difference
1000
2000
3000
4000
5000
-500
-250
0
250
500


Statistical power comparison
0
0.05
0.1
Statistical Power
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1000
2000
3000
4000
5000
Sample size per group
Statistical Power of t-test on CUPED-adjusted metric increases along with sample size / effect size
Type 1 error rate under 5%
Statistical Power of t-test on original metric barely increased
test
t.test
t.test_cuped

Factors that affect CUPED

The higher the better

$$Var(y^*) = Var(y) \cdot (1 - \rho_{x,y}^2)$$



Coverage

$$\frac{A}{A+B}$$

A Units enrolled with pre-experiment data

B Units enrolled without pre-experiment data

Choosing other covariate

As long as the treatment does NOT affect the covariate

$$E(X_c) = E(X_t)$$


Microsoft
CUPED
"Let's see who you really are!"

Well…kinda

CUPED

```
theta = cov(post, pre)/var(pre)
post_cuped = post - theta*(pre - mean(pre))
```

```
t.test(post_cuped ~ cohort)
```

```
        Welch Two Sample t-test

data:  post_sim_cuped by cohort
t = -2.6958, df = 8838.3, p-value = 0.007036
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -219.06128  -34.60611
sample estimates:
mean in group ctrl  mean in group trt
          3468.314            3595.148
```

Difference = 126.834

ANCOVA

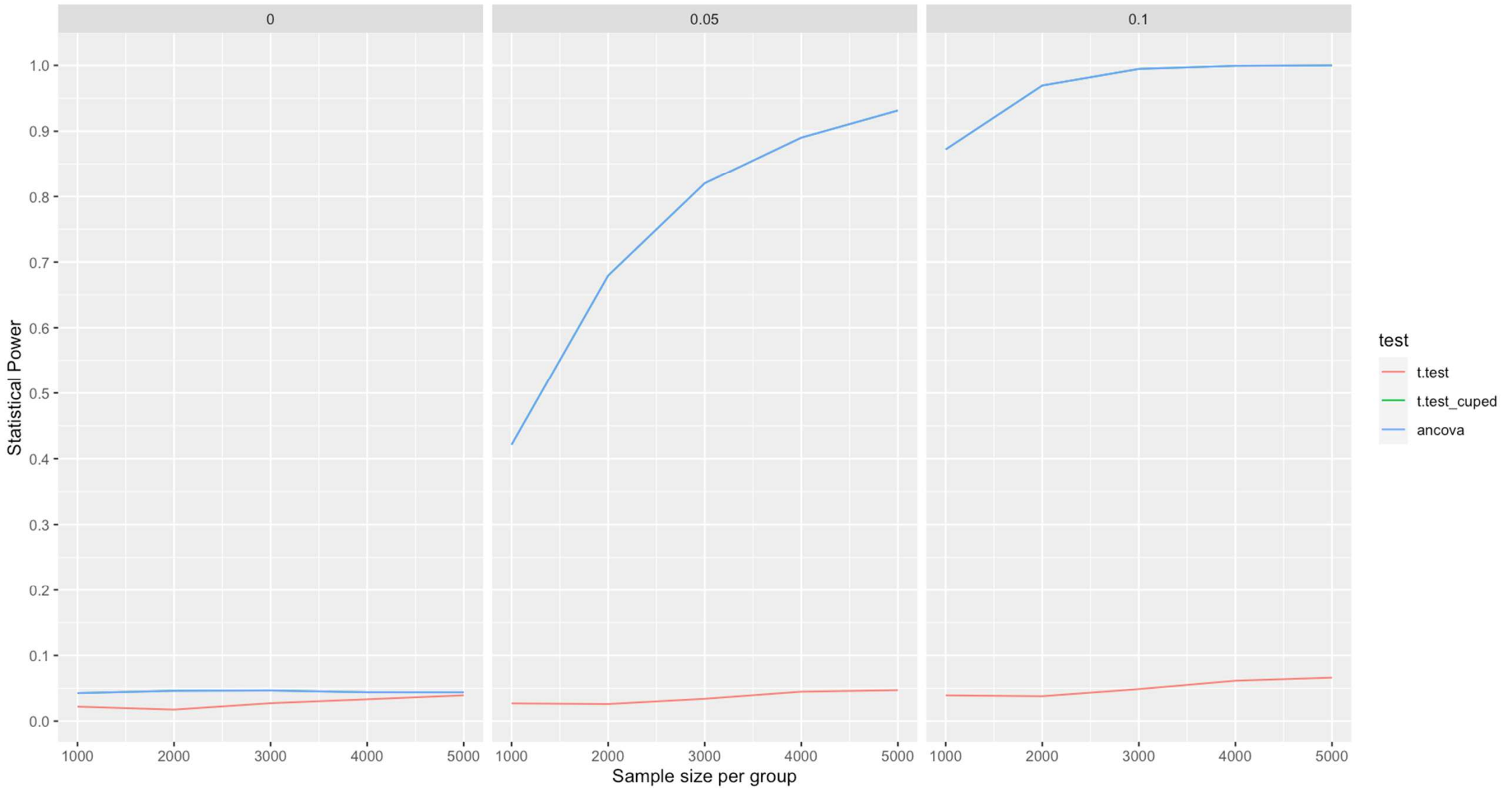
```
lm(post ~ pre + cohort)
```

```
Call:
lm(formula = post_sim ~ pre + cohort, data = rd_cuped_sample)

Residuals:
    Min      1Q  Median      3Q     Max
-120404    -150       8     147   89628

Coefficients:
              Estimate  Std. Error  t value Pr(>|t|)
(Intercept) -93.4427750  33.3436476   -2.802  0.00508 **
pre           1.1623613   0.0007037 1651.698  < 2e-16 ***
cohorttrt   126.8342718  47.0517726    2.696  0.00704 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2353 on 9997 degrees of freedom
Multiple R-squared:  0.9963,    Adjusted R-squared:  0.9963
F-statistic: 1.364e+06 on 2 and 9997 DF,  p-value: < 2.2e-16
```


Statistical power comparison
0
0.05
0.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Statistical Power
1000
2000
3000
4000
5000
Sample size per group
test
t.test
t.test_cuped
ancova

Different way of solving problem

CUPED is simple to implement and performant from a coding perspective



Thank you!