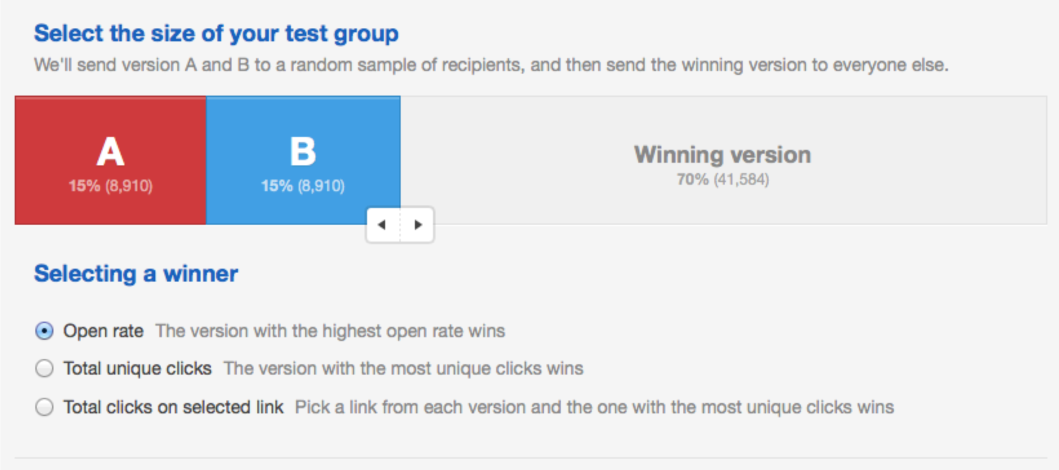
Test & Roll: Profit Maximizing Marketing Experiments

Elea McDonnell Feit (Drexel) & Ron Berman (Wharton)

Paper: `https://arxiv.org/abs/1811.00457`
Code: `https://github.com/eleafeit/testandroll`

The test & roll problem

Email test & roll setup



Source: Zapier.com

Standard analysis: difference-in-means hypothesis test

$$\bar{y}_1 - \bar{y}_2 \geq z_{1-\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$\bar{y}_1, \bar{y}_2$: average response
s_1, s_2: s.d. of response
n_1, n_2: customers in test groups
α: chance of false positive

If the inequality holds the null hypothesis is rejected and the difference in mean response between groups 1 and 2 is declared significant.

Standard sample size recommendation

$$n_{HT} = n_1 = n_2 \approx 2(z_{1-\alpha/2+z_\beta})^2 \left(\frac{2s^2}{d^2}\right)$$

s: s.d. of response

d: difference to detect

α: chance of false positive (significance)

β: chance of a false negative if the true difference is d

Limitations of hypothesis testing

False positives do not reduce profit in this setting. Why control them?

Limitations of hypothesis testing

False positives do not reduce profit in this setting. Why control them?

What if recommended sample size is larger than available population?

Limitations of hypothesis testing

False positives do not reduce profit in this setting. Why control them?

What if recommended sample size is larger than available population?

Which treatment should be deployed if the difference is non-significant?

Limitations of hypothesis testing

False positives do not reduce profit in this setting. Why control them?

What if recommended sample size is larger than available population?

Which treatment should be deployed if the difference is non-significant?

Can't rationalize unequal test group sizes.

Limitations of hypothesis testing

False positives do not reduce profit in this setting. Why control them?

What if recommended sample size is larger than available population?

Which treatment should be deployed if the difference is non-significant?

Can't rationalize unequal test group sizes.

Rate of false positives can be inflated by ongoing monitoring of tests.

Profit-maximizing A/B tests

Profit for a test & roll

The goal of a test & roll experiment with a finite population N is to maximize total profit earned, which is the profit earned in the test and deploy stages:

$$\text{Profit}_{\text{Test}} + \text{Profit}_{\text{Deploy}}$$

Profit for a test & roll

The goal of a test & roll experiment with a finite population N is to maximize total profit earned, which is the profit earned in the test and deploy stages:

$$\text{Profit}_{\text{Test}} + \text{Profit}_{\text{Deploy}}$$

Under this framework, we can set sample size n_1 and n_2 to maximize profit:

$$(n_1^*, n_2^*) = \underset{n_1, n_2}{argmax}\, E[\text{Profit}_{\text{Test}} + \text{Profit}_{\text{Deploy}}]$$

This involves a trade-off between the opportunity cost of the test, where some people get a suboptimal treatment, and the likelihood of choosing the better treatment to deploy over the limited deployment population.

Symmetric Normal-Normal model

Distribution of profit per customer:

$$Y_1 \sim N(m_1, s^2),\, Y_2 \sim N(m_2, s^2)$$

Priors:

$$m_1, m_2 \sim N(\mu, \sigma^2),\, s \text{ known}$$

Optimal decision rule:

Choose the treatment with the greater posterior mean.
Under symmetric priors this means choose treatment j if $\bar{y}_j$ is larger.

Expected profit for Normal-Normal

Solving for *a priori* expected profit:

$$E[\text{Profit}_{\text{Test}} + \text{Profit}_{\text{Deploy}}] = \mu N + (N - n_1 - n_2)\left(\frac{\sqrt{2}\sigma^2}{\sqrt{\pi}\sqrt{\frac{n_1+n_2}{n_1 n_2}s^2 + 2\sigma^2}}\right)$$

Term in parentheses is the increased expected profit due to choosing the correct treatment to deploy. When n_1 and n_2 are larger, the increase in expected profit is greater, but is earned for fewer customers.

Profit-maximizing sample size for Normal-Normal

Solving for sample sizes to maximize expected profit:

$$n_1^* = n_2^* = \sqrt{\frac{N}{4}\left(\frac{s}{\sigma}\right)^2 + \left(\frac{3}{4}\left(\frac{s}{\sigma}\right)^2\right)^2} - \frac{3}{4}\left(\frac{s}{\sigma}\right)^2 \le \sqrt{N}\frac{s}{2\sigma}$$

Profit-maximizing sample size for Normal-Normal

Solving for sample sizes to maximize expected profit:

$$n_1^* = n_2^* = \sqrt{\frac{N}{4}\left(\frac{s}{\sigma}\right)^2 + \left(\frac{3}{4}\left(\frac{s}{\sigma}\right)^2\right)^2} - \frac{3}{4}\left(\frac{s}{\sigma}\right)^2 \leq \sqrt{N}\frac{s}{2\sigma}$$

Higher when the population size (N) is larger

Profit-maximizing sample size for Normal-Normal

Solving for sample sizes to maximize expected profit:

$$n_1^* = n_2^* = \sqrt{\frac{N}{4}\left(\frac{s}{\sigma}\right)^2 + \left(\frac{3}{4}\left(\frac{s}{\sigma}\right)^2\right)^2} - \frac{3}{4}\left(\frac{s}{\sigma}\right)^2 \leq \sqrt{N}\frac{s}{2\sigma}$$

Higher when the population size (N) is larger

Grows sub-linearly with the standard deviation s (versus s^2)

Profit-maximizing sample size for Normal-Normal

Solving for sample sizes to maximize expected profit:

$$n_1^* = n_2^* = \sqrt{\frac{N}{4}\left(\frac{s}{\sigma}\right)^2 + \left(\frac{3}{4}\left(\frac{s}{\sigma}\right)^2\right)^2} - \frac{3}{4}\left(\frac{s}{\sigma}\right)^2 \leq \sqrt{N}\frac{s}{2\sigma}$$

Higher when the population size (N) is larger

Grows sub-linearly with the standard deviation s (versus s^2)

Smaller when greater difference in performance between treatments (σ)

Regret for Normal-Normal

$$E[\text{Profit}|\text{Perfect Information}] - E[\text{Profit}_{\text{Test}} + \text{Profit}_{\text{Deploy}}] =$$

$$N\frac{\sigma}{\sqrt{\pi}}\left(1 - \frac{\sigma}{\sqrt{\sigma^2 + \frac{s^2}{n*}}}\right) + \frac{2n^*\sigma^2}{\sqrt{\pi}\sqrt{\sigma^2 + \frac{s^2}{n*}}} \leq O(\sqrt{N})$$

Regret for Normal-Normal

$$E[\text{Profit}|\text{Perfect Information}] - E[\text{Profit}_{\text{Test}} + \text{Profit}_{\text{Deploy}}] =$$

$$N\frac{\sigma}{\sqrt{\pi}}\left(1 - \frac{\sigma}{\sqrt{\sigma^2 + \frac{s^2}{n*}}}\right) + \frac{2n^*\sigma^2}{\sqrt{\pi}\sqrt{\sigma^2 + \frac{s^2}{n*}}} \leq O(\sqrt{N})$$

Test & roll compares favorably to a multi-armed bandit with Thompson sampling which also has regret $O(\sqrt{N})$.

Application

Profit-maximizing sample size

Distribution of profit per customer:

$$Y_1 \sim N(m_1, s^2),\ Y_2 \sim N(m_2, s^2)$$

Priors:

$$m_1, m_2 \sim N(\mu, \sigma^2),\ s \text{ known}$$

Profit-maximizing sample size:

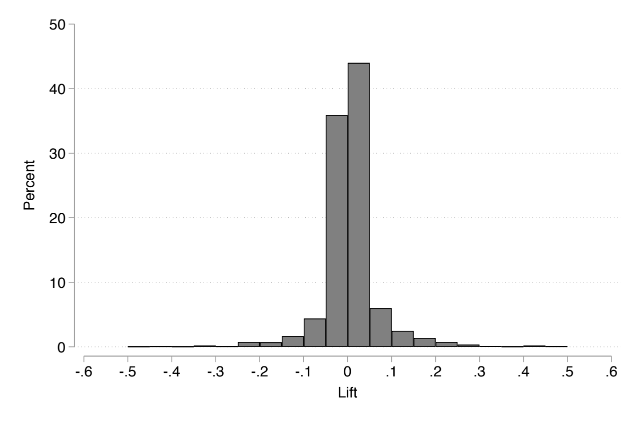
$$n_1^* = n_2^* = \sqrt{\frac{N}{4}\left(\frac{s}{\sigma}\right)^2 + \left(\frac{3}{4}\left(\frac{s}{\sigma}\right)^2\right)^2} - \frac{3}{4}\left(\frac{s}{\sigma}\right)^2 \leq \sqrt{N}\frac{s}{2\sigma}$$

Previous A/B website tests

2,101 website A/B tests

Each user i in each test k is randomly assigned to treatment $j \in 1, 2$ and we observe whether or not they "clicked"

Treatments are exchangable



Meta-analysis of website tests

Model

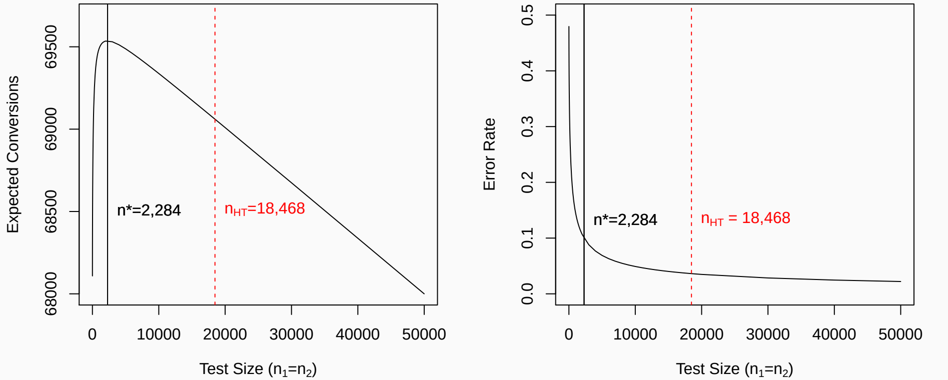
$$\text{response: } y_{ijk} \sim \mathcal{N}(m_{jk}, s)$$
$$\text{treatment mean: } m_{jk} \sim \mathcal{N}(t_k, \sigma)$$
$$\text{test mean : } t_k \sim \mathcal{N}(\mu, \omega)$$

Posterior

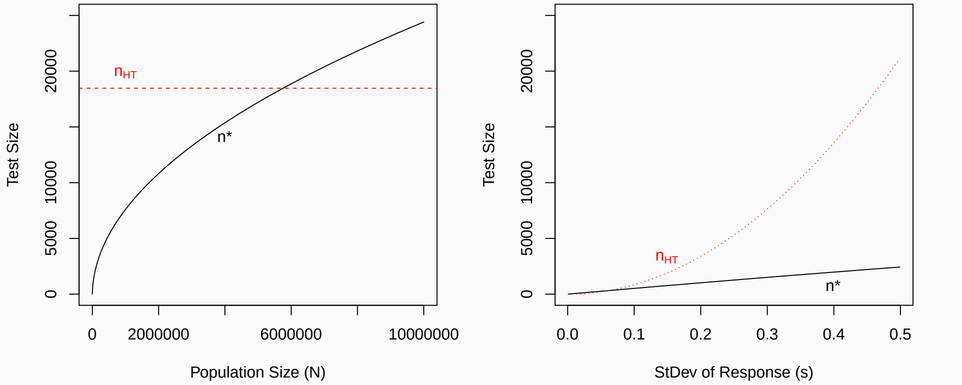
	mean	sd	2.5%-tile	97.5%tile
μ	0.676	0.004	0.667	0.685
σ	0.030	0.001	0.029	0.031
ω	0.199	0.003	0.193	0.206

Expected profit (conversions)

Assuming $\mu = 0.676$ and $\sigma = 0.030$ and total population of $N = 100,000$ the optimal sample size is $n* = 2,284$ in each group.



Profit-maximizing test size varies with N and s



Profit for alternative test methods

	n_1	n_2	Expected Conversions Test	Expected Conversions Roll	Expected Conversions Overall	Regret	Roll Error
No Test (Random)	-	-	-	-	68,000	2.43%	50.0%
Hypothesis Test	18,468	18,468	25,116	43,944	69,060	0.91%	3.6%
Test & Roll	2,284	2,284	3,106	66,430	69,536	0.22%	10.0%
Thompson Sampling	-	-	-	-	69,637	0.08%	-
Perfect Information	-	-	-	-	69,693	0%	-

Alternative models (details in paper)

Normal-Normal with asymmetric priors

- Incumbent/challenger tests
- Media holdout tests (Catalog example in paper)
- Pricing or discount tests

Alternative models (details in paper)

Normal-Normal with asymmetric priors

- Incumbent/challenger tests
- Media holdout tests (Catalog example in paper)
- Pricing or discount tests

Beta-Binomial

- Important to consider probability of ties when n_1 and n_2 are small.

Alternative models (details in paper)

Normal-Normal with asymmetric priors

- Incumbent/challenger tests
- Media holdout tests (Catalog example in paper)
- Pricing or discount tests

Beta-Binomial

- Important to consider probability of ties when n_1 and n_2 are small.

Both require numerical optimization to find the sample size.

Conclusion

Benefits of profit-maximizing experiments

Sample sizes are substantially reduced versus standard recommendations, especially when response is noisy.

Sample sizes are proportional to the total available population, providing a rational recommendation when total population is small.

Analysis is straightforward and intuitive (often simply “pick the winner”).

Unequal group sizes are rationalized.

Thanks!

Elea McDonnell Feit
Assistant Professor of Marketing
Drexel University
eleafeit@gmail.com
@eleafeit

Paper: `https://arxiv.org/abs/1811.00457`
Code: `https://github.com/eleafeit/testandroll`

Backup

Error rate for normal-normal

The error rate in deployment is:

$$E[I(\bar{y}_1 > \bar{y}_2)|m_1 < m_2)] = E[I(\bar{y}_1 < \bar{y}_2)|m_1 > m_2)] =$$
$$\frac{1}{4} - \frac{1}{2\pi} \arctan\left(\frac{\sqrt{2}\sigma}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}\right)$$

However, unlike hypothesis testing, the error rate is determined by making the profit-maximizing trade-off between the opportunity cost of the test and risk of an incorrect deployment.

Distribution of regret relative to Thompson sampling