When Your A/B Test Hits the MDE: Cracking the p-value Puzzle

Quang Tri Le
LinkedIn: lequangtri Telegram: @lequangtri
GitHub: leqtr/Experimentation-DeepDives

1 Introduction

Recently, I came across a classic brain teaser from A/B-testing guru Ron Kohavi:

> "You design an A/B test to detect a relative effect of at least 5% (MDE) to your conversion rate using industry recommended parameters: alpha=0.05 (5% type-I error) and 80% power (20% type-II error). When the experiment finishes and reaches the planned number of users per above, you see that the treatment effect is exactly 5%.
> *What is the p-value?*"

Naturally, I wanted to test my intuition and solve the problem in my mind. **Spoiler**: I failed. My first thought was: do we have enough information to compute a p-value? Should we know the base conversion?
Armed with nothing but pen and paper, I was determined to find the answer. **Another spoiler**: Strictly speaking, the problem omits a couple of technical details needed for an exact p-value, yet it still provides enough context to derive a very close approximation.

Before you continue, I encourage you to pause and try solving this puzzle yourself. It's a great exercise to solidify your understanding of core statistical concepts behind A/B testing.

In the sections that follow, I will:

1. Break down the analytical solution step by step
2. Use visualizations of the core concepts to build intuition
3. Discuss the assumptions and limitations of the solution

2 Statistical Setup

We are running an A/B test on binary conversion rate which corresponds to testing difference in means of two groups (consisting of 0 and 1) for significance.

$$H_0 : \mu_1 - \mu_0 = 0$$
$$H_1 : \mu_1 - \mu_0 > 0$$

I will use a one-sided alternative to simplify visualizations, but the idea remains the same for a two-sided alternative.

Let X denote conversion of a user in the control group, $X \sim Bernoulli(p_0)$, and Y - conversion of a user in pilot group, $Y \sim Bernoulli(p_1)$, where $p_1 = 1.05p_0$ (5% relative effect designed). Then conversion rates in groups can be represented as sample means.

$$\bar{X} = \frac{X_1 + ... + X_{n_x}}{n_x} \overset{CLT}{\sim} \mathcal{N}(p_0, \frac{p_0(1-p_0)}{n_x})$$
$$\bar{Y} = \frac{Y_1 + ... + Y_{n_y}}{n_y} \overset{CLT}{\sim} \mathcal{N}(p_1, \frac{p_1(1-p_1)}{n_y})$$

We assume the standard A/B-test conditions are satisfied: observations within each group are independent and identically distributed, the two groups are independent of one another, and the sample sizes are sufficiently large for the Central Limit Theorem to yield approximate normality of the sample means. Building the test statistic - difference in means:

$$\Delta = \bar{Y} - \bar{X} \overset{CLT}{\sim} \mathcal{N}(p_1 - p_0, \frac{p_1(1-p_1)}{n_y} + \frac{p_0(1-p_0)}{n_x})$$

When outcomes are binary, the group means coincide with conversion proportions, so

$$\mu_1 - \mu_0 = p_1 - p_0.$$

Thus a 5% relative lift yields an absolute MDE of

$$MDE = p_1 - p_0 = 1.05\,p_0 - p_0 = 0.05\,p_0.$$

3 Statistic's Distribution under H_0

Under H_0,

$$\Delta \overset{H_0}{\sim} \mathcal{N}(0, \frac{p_0(1-p_0)}{n_y} + \frac{p_0(1-p_0)}{n_x}) = \mathcal{N}(0, (SE_0)^2)$$

Let's visualize the distribution of our test statistic under H_0, show the critical cutoff, mark the observed uplift (the MDE), and illustrate the resulting p-value:

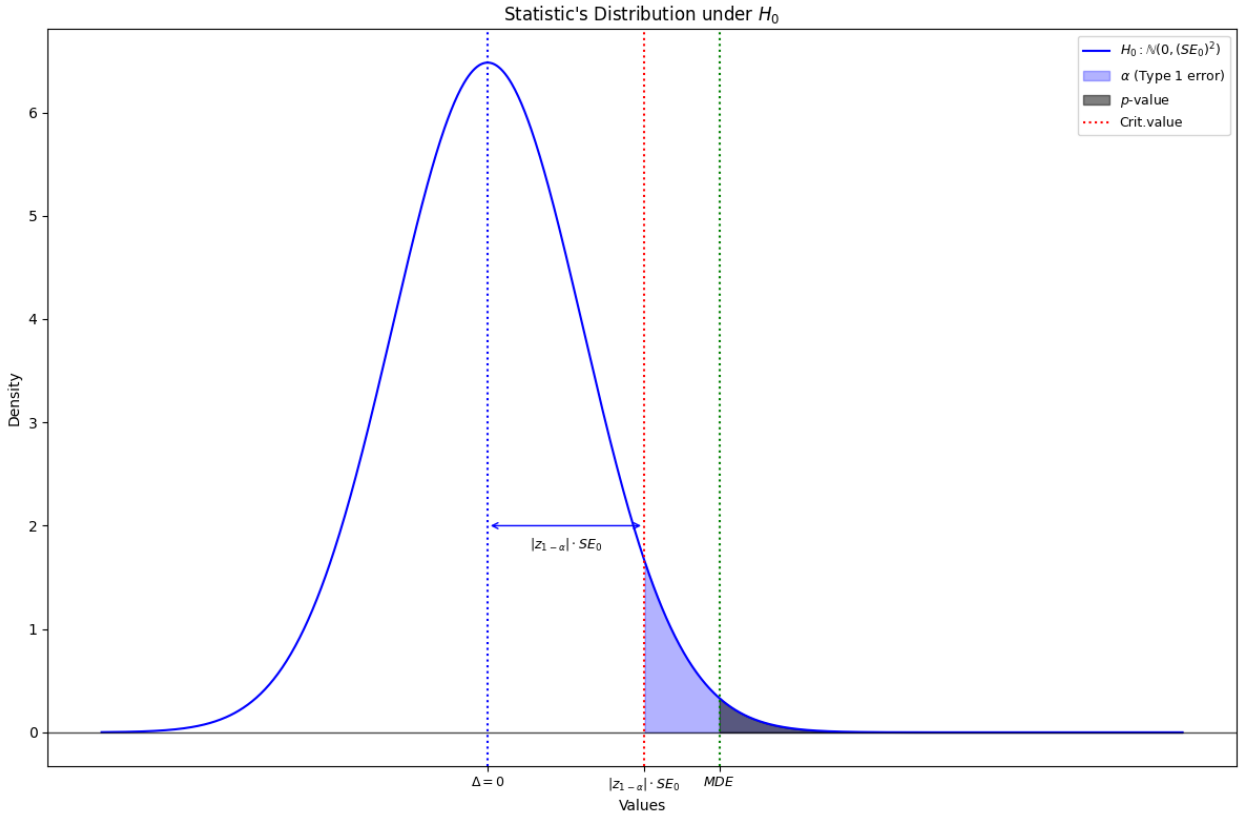

Figure 1: Null distribution (H_0); shaded area = α=0.05; dark tail = observed p-value.

Since I picked one-sided test, **significance bound is** $|z_{1-\alpha}| \cdot SE_0$, where $z_{1-\alpha}$ - critical value of the standard normal distribution and SE_0 - standard error of the difference in means statistic under H_0. If observed uplift is more extreme than scaled critical value, we reject H_0. This rejection would be a Type 1 error, if H_0 is was actually true (the difference = 0).

4 Statistic's Distribution under H_0 and H_1

Additionally, we know that we have designed the experiment such that under the alternative hypothesis (H_1) our statistic also has approximately normal distribution centered around our MDE:

$$\Delta \overset{H_1}{\sim} \mathcal{N}(MDE, \frac{p_1(1-p_1)}{n_y} + \frac{p_0(1-p_0)}{n_x}) = \mathcal{N}(MDE, (SE_1)^2)$$

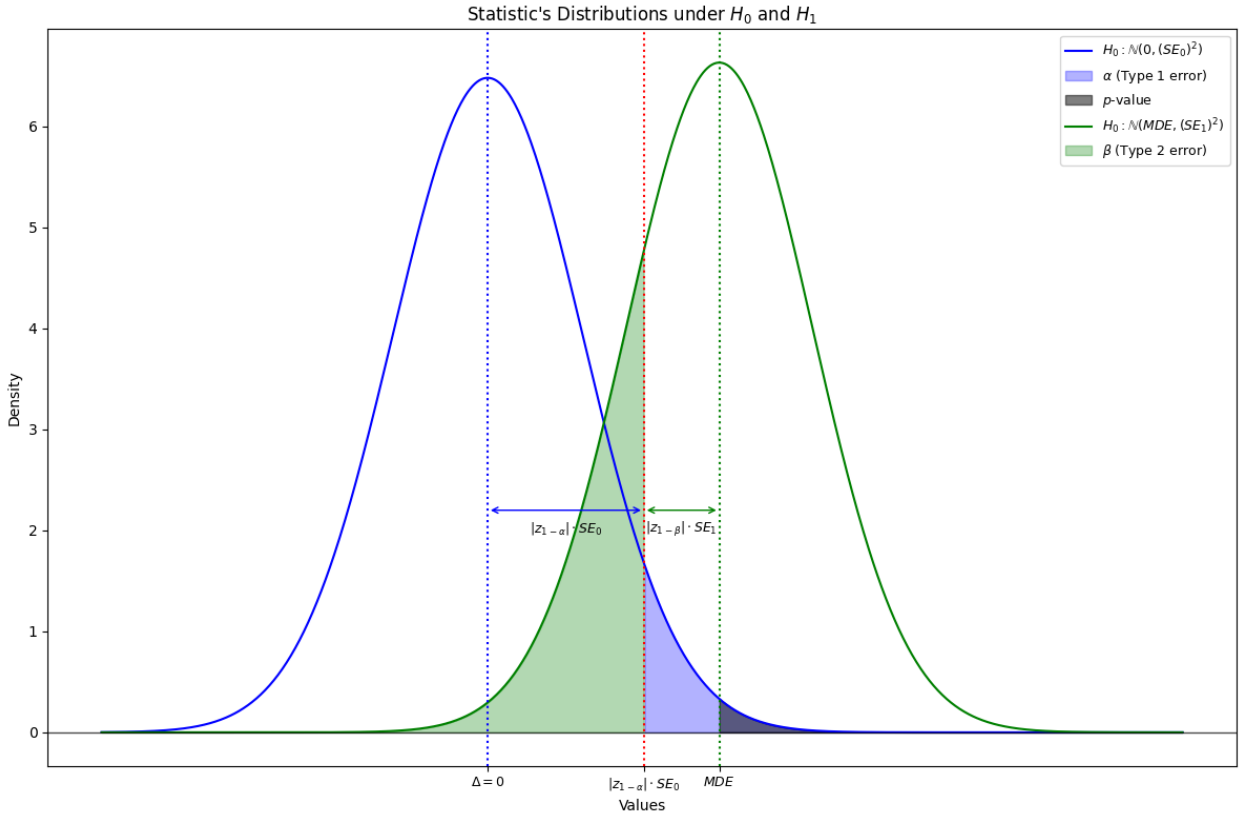

Figure 2: Null (blue) vs. alternative (green) distributions, illustrating α and β.

Not rejecting H_0, when it is False (Type 2 error) corresponds to green shaded area, $\beta = 0.2$. The distance between H_1 distribution mean and critical bound equals to $|z_{1-\beta}| * SE_1$. Combining together distances from critical bound to H_0 mean and H_1 mean:

$$MDE = |z_{1-\alpha}| \cdot SE_0 + |z_{1-\beta}| \cdot SE_1$$

5 Standardized Statistic's Distribution under H_0 and H_1

To calculate p-value we standardize the test statistic by subtracting mean and dividing by standard error of H_0 distribution:

$$\Delta_{st} = \frac{\Delta - 0}{SE_0}$$
$$\Delta_{st} \overset{H_0}{\sim} \mathcal{N}(0, 1^2)$$
$$\Delta_{st} \overset{H_1}{\sim} \mathcal{N}(\frac{MDE}{SE_0}, (\frac{SE_1}{SE_0})^2)$$

After standardizing by SE_0, the null distribution becomes $\mathcal{N}(0, 1)$, but the alternative still has variance $(SE_1/SE_0)^2 \neq 1$.

Standardized observed uplift and p-value

$$MDE_{st} = \frac{MDE}{SE_0} = |z_{1-\alpha}| + |z_{1-\beta}| \cdot \frac{SE_1}{SE_0}$$
$$p - value = 1 - \Phi(MDE_{st}) = 1 - \Phi(|z_{1-\alpha}| + |z_{1-\beta}| \cdot \frac{SE_1}{SE_0})$$

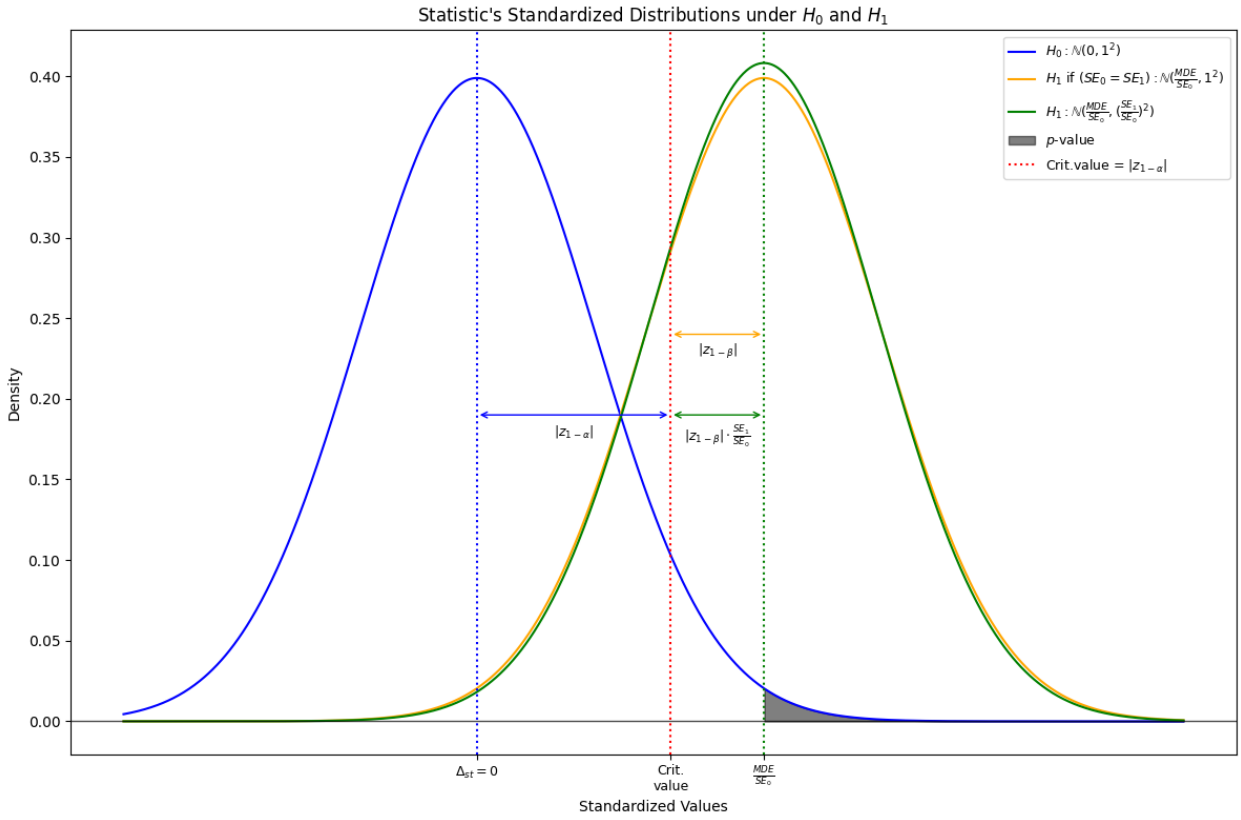

Figure 3: Standardized distributions under H_0 and H_1.

We have arrived at the final part of the solution. $|z_{1-\alpha}|$ and $|z_{1-\beta}|$ are constants that we can look up in the normal distribution table. The value of MDE_{st} and, hence, p-value depends on $\frac{SE_1}{SE_0}$:

- if $\frac{SE_1}{SE_0} \approx 1$ (yellow distribution in Figure 3), $p - value \approx 1 - \Phi(|z_{1-\alpha}| + |z_{1-\beta}|)$
- **if $\frac{SE_1}{SE_0} \not\approx 1$, then we can't obtain a good estimate of the $p - value$ without knowing conversion baseline (to calculate SE_0 and SE_1)**

6 Estimating $\frac{SE_1}{SE_0}$ and $p - value$

Although we can't calculate a precise $p - value$ without knowing conversion baseline, we can bound $\frac{SE_1}{SE_0}$ and the resulting $p - value$ as a function of p_0. For simplicity of calculations, let's assume 50/50 split (most common split in practice), i.e. $n_x = n_y = n$:

$$(1)\quad (\frac{SE_1}{SE_0})^2 = \frac{\frac{1.05p_0(1-1.05p_0)}{n_y} + \frac{p_0(1-p_0)}{n_x}}{\frac{p_0(1-p_0)}{n_y} + \frac{p_0(1-p_0)}{n_x}} = \frac{\frac{1.05(1-1.05p_0)}{1-p_0} + 1}{2}$$

$$\frac{d}{dp}(\frac{1-1.05p}{1-p}) = \frac{-1.05(1-p) - (1-1.05p)(-1)}{(1-p)^2} = \frac{-0.05}{(1-p)^2} < 0$$

$$(2)\quad p - value = 1 - \Phi(MDE_{st}) = 1 - \Phi(|z_{1-\alpha}| + |z_{1-\beta}| \cdot \frac{SE_1}{SE_0})$$

$$|z_{1-\alpha}| + |z_{1-\beta}| \cdot \frac{SE_1}{SE_0} \downarrow\rightarrow \Phi(.) \downarrow\rightarrow p - value \uparrow$$

$$(1) + (2) => p_0 \uparrow\rightarrow \frac{SE_1}{SE_0} \downarrow\rightarrow p - value \uparrow$$

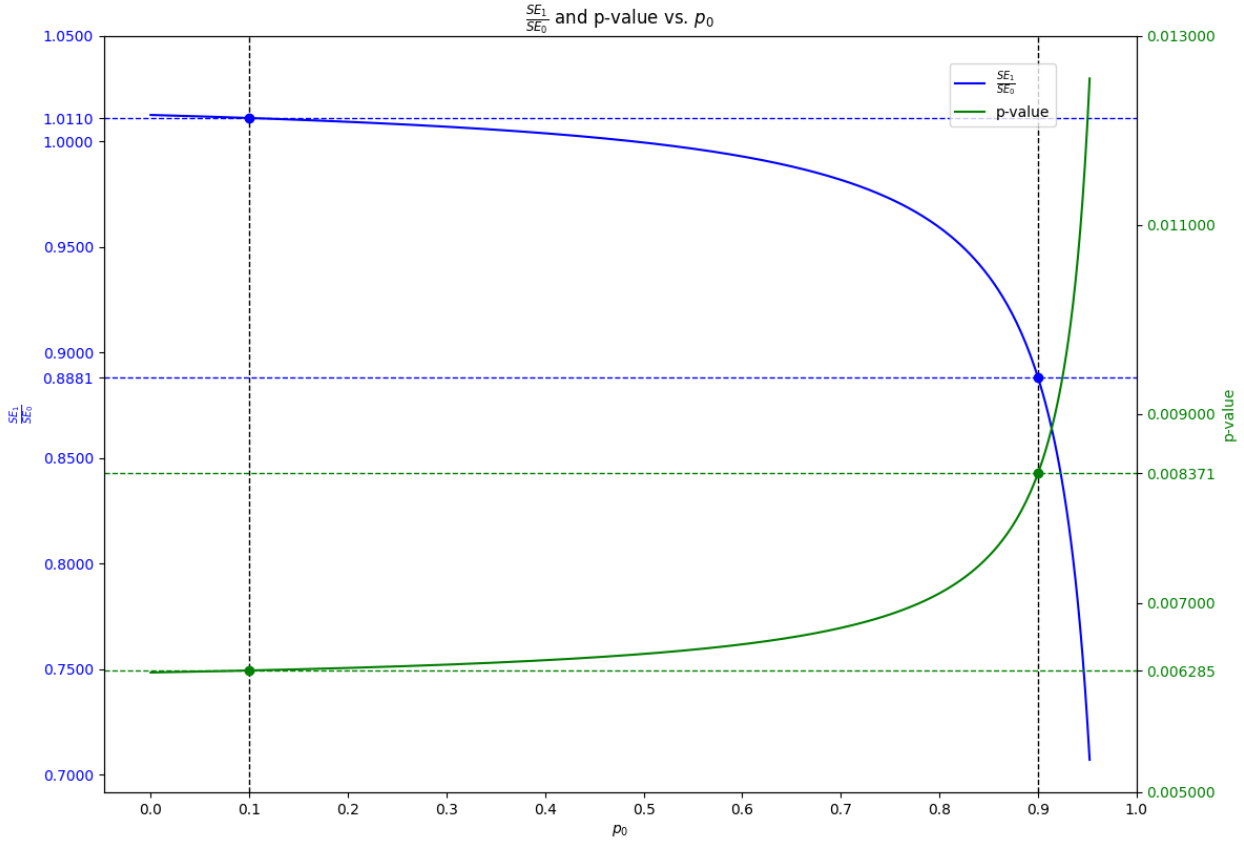

Figure 4: Dependence of SE-ratio and p-value on baseline conversion rate

- As p_0 increases from 0 to $1/1.05 \approx 0.9524$, the ratio $\frac{SE_1}{SE_0}$ decreases from about 1.025 to 0.707.
- Over the same range, the p-value increases from approximately 0.0061 to 0.0125.

In practice, many binary conversion metrics (n-day retention, activation rate, survey responses, etc.) lie within a 10%–90% band. Hence, we focus on $p_0 \in [0.1, 0.9]$ for practical insights.

7 Final answer

- Under the assumption $\frac{SE_1}{SE_0} = 1$, $p - value = 1 - \Phi(|z_{1-\alpha}| + |z_{1-\beta}|) \approx 0.0065$
- For a baseline conversion in the practical range $[10\%, 90\%]$, $p - value$ varies from approximately 0.0063 to 0.0084, which remains well below 0.05, so the result would be significant in every case.
- Even at extreme baseline rates (below 10% or above 90%), the result remains statistically significant, though the p-value can vary more dramatically.

We've used simplifying assumptions and specific parameter values for clarity. Consider how the results would change if you:

- Use a two-sided instead of a one-sided alternative.
- Adopt an unequal traffic split rather than 50/50.
- Choose a different MDE (larger or smaller).
- Choose a different Power level (especially ¡ 50%).

8 Bonus: beyond binary metrics

The same analytical flow applies to absolute metrics, not just binary conversions, if you are working with the difference in means statistic (Z-test, T-test, Welch's test). However, when testing non-binary metrics, there are a couple of pitfalls you should watch out for:

- **Variance shifts:** Unlike binary conversion rates, absolute metrics (revenue, session length, etc.) can exhibit large variance changes between control and treatment. A substantial variance shift alters the standard error under H_1 and can flip your test's significance.
- **Heavy tails & outliers:** Some absolute metrics have heavy-tailed distributions and/or contain significant share of extreme outliers, which can break CLT-based normal approximation for sample means. As a best practice, run simulations on historical data to confirm that your chosen inference method is appropriate and reliable.

9 Conclusion

Here are the key takeaways I hope you'll find useful:

- Powering for a specific effect and then observing exactly that uplift yields a p-value well below your nominal significance level, giving you very strong evidence.
- Visualizing the null and alternative distributions makes it clear how your inputs (MDE, α, power) determine the p-value.
- When some details are missing, clearly state and justify your assumptions, then derive approximate bounds to deliver a structured, case-by-case answer.

10 Code for Generating the Figures

You can view and run the Python code used to generate all figures in this article in my GitHub: leqtr/Experimentation-DeepDives

11 References

- Ron Kohavi, "Test your intuition on p-values?"