CS109 The Probability Challenge

HAEJUNG JUNG

2026-08-10

The Pitfalls of A/B Testing — "Why Ad A/B Tests Lie"

Note on Data: All numerical metrics and figures referenced below are reconstructed illustrative examples based on real-world campaign dynamics.

LLM as a Tool: This report was prepared with the help of Gemini and translated into English.

1. Motivation & Background

In performance marketing, A/B testing is the cornerstone of decision-making, yet statistical pitfalls frequently lead to misguided conclusions. Prior to studying computer science, I managed digital performance marketing campaigns for a domestic cosmetics brand. During one campaign, we conducted an A/B test comparing a "Functional Appeal" creative (highlighting active ingredients) against an "Emotional Appeal" creative (targeting consumer pain points).

In the early stages of the test (impressions $N < 500$), the emotionally provocative creative achieved an observed Click-Through Rate (CTR) of over 5%, seemingly crushing the functional creative (1%). Standard industry practice strongly encourages immediately reallocating the budget to such early winners. However, as data accumulated to $N = 10{,}000$, this early outperformance proved to be mere noise driven by the high variance of a small sample. Ultimately, the functional creative converged to its true value ($\theta_A = 0.022$) and emerged as the clear winner. This experience highlighted a critical problem: statistical illusions that induce premature decision-making lead directly to massive marketing budget waste.

2. Pain Points of Frequentist A/B Testing

The prevailing p-value-based frequentist A/B testing framework suffers from several fatal flaws in practice:

- **The Peeking Problem (Type I Error Inflation):** Marketers constantly monitor real-time dashboards and often stop tests the moment $p < 0.05$. This practice is equivalent to continuous multiple hypothesis testing. It inflates the False Positive rate (Type I Error) from the nominal 5% to over 30%, causing teams to frequently declare winners when no true difference exists (H_0).
- **Misuse of the Central Limit Theorem (CLT):** In the early stages ($N < 500$), the sample size is insufficient for the normal approximation of a Binomial distribution to hold. Yet, standard z-tests are routinely applied, outputting distorted confidence metrics.
- **The Zero-Prior Trap:** Despite strong domain knowledge dictating that average beauty industry CTRs range from 1.5% to 2.5%, frequentist inference relies entirely on the Maximum Likelihood Estimate (MLE, $\hat{\theta} = \frac{k}{n}$). This causes the model to severely overfit to early outliers (e.g., 5%).

3. The Bayesian Solution: Beta-Binomial Model

This project resolves the structural flaws of frequentist A/B testing by applying the conjugate prior and Bayesian inference frameworks taught in CS109.

A. Beta-Binomial Conjugate Model & Prior Setup

The total number of clicks k out of n impressions (Bernoulli trials) follows a Binomial likelihood. By assigning a Beta distribution as the prior for the CTR parameter θ, the posterior distribution updates analytically:

$$\theta \mid \text{data} \sim \text{Beta}(\alpha + k,\ \beta + n - k)$$

Reflecting the beauty industry average CTR (2%), we establish an Informative Prior of $\text{Beta}(2, 98)$. Mathematically, this is equivalent to injecting $N_0 = 100$ pseudo-counts (prior observations) into the model before the test even begins.

B. Shrinkage Effect (Regularization & Noise Defense)

Even if an early observation yields 15 clicks out of $N = 300$ (an MLE of 5%), the Maximum A Posteriori (MAP) estimate is regularized by the prior. This "shrinkage effect" pulls the estimate back to a realistic range:

$$\hat{\theta}_{\text{MAP}} = \frac{k + \alpha - 1}{n + \alpha + \beta - 2} = \frac{15 + 1}{300 + 98} \approx 4.02\%$$

As data (n) approaches infinity, the prior's influence diminishes, and the observed data (likelihood) naturally dominates the posterior.

C. Direct Probability Computation via Monte Carlo Sampling

Instead of relying on asymptotic assumptions, we draw 10,000 samples from the posterior distributions of both creatives to directly compute $P(\theta_B > \theta_A)$. This delivers a highly intuitive, business-friendly metric—e.g., *"There is an 82% probability that Creative A outperforms B."*

4. Expected Loss Stopping Rule

To determine precisely *when* to stop an experiment, we replace arbitrary variance reduction or p-value thresholds with a Bayesian Expected Loss framework. The potential loss incurred by choosing Creative A is defined as:

$$\mathbb{E}[\text{Loss}_A] = \iint \max(\theta_B - \theta_A,\, 0) \cdot P(\theta_A \mid \text{data}_A)\, P(\theta_B \mid \text{data}_B)\, d\theta_A\, d\theta_B$$

The algorithm forces the test to continue until the Expected Loss drops below a business-acceptable threshold ϵ (e.g., a CTR drop of $0.05\%p$, $\epsilon = 0.0005$), completely neutralizing the risk of premature termination.

5. Simulation & Visualization Results

Using Python (NumPy, SciPy), we conducted $M = 10{,}000$ Monte Carlo simulations to derive three visual proofs:

1. **Fig 1. Posterior Convergence:** As N grows from 300 to 10,000, the initially wide posterior distributions narrow sharply around the true values (2.2% and 2.0%), resolving uncertainty ($\sigma^2 \propto \frac{1}{N}$).
2. **Fig 2. Optimal Decision via Expected Loss:** We plotted the narrowing 95% Credible Interval alongside the Expected Loss curve, quantitatively identifying the safe stopping point where the loss curve dips below ϵ.
3. **Fig 3. The Peeking Trap Simulation:** Under the null hypothesis (H_0), continuous p-value peeking caused the Cumulative False Positive Rate to explode to $\sim 35\%$. Conversely, the Bayesian Expected Loss rule successfully constrained the alpha error, remaining flat and stable.

6. Expected Business Impact

This framework dismantles the blind reliance on p-values. By formally integrating domain knowledge (Priors) and quantifying uncertainty (Expected Loss), it establishes a genuinely data-driven culture. Ultimately, it prevents the fatal mistake of shifting thousands of dollars toward underperforming assets due to statistical noise.