A/B Testing Project — Cookie Cats Mobile Game

This project demonstrates an end-to-end A/B testing workflow using the Cookie Cats mobile gaming dataset while simultaneously validating the custom A/B testing automation functions I developed. The goal was two-fold:

1. Analyze the real experimental question: Does moving the level gate from 30 to 40 improve engagement and retention?
2. Test and verify my A/B testing pipeline:
    - SRM check
    - A/A test
    - Proportion tests
    - Mean tests
    - Delta method for ratio metrics
    - Simpson's paradox checks
    - Combined results table
    - Effect size labelling
    - Power analysis

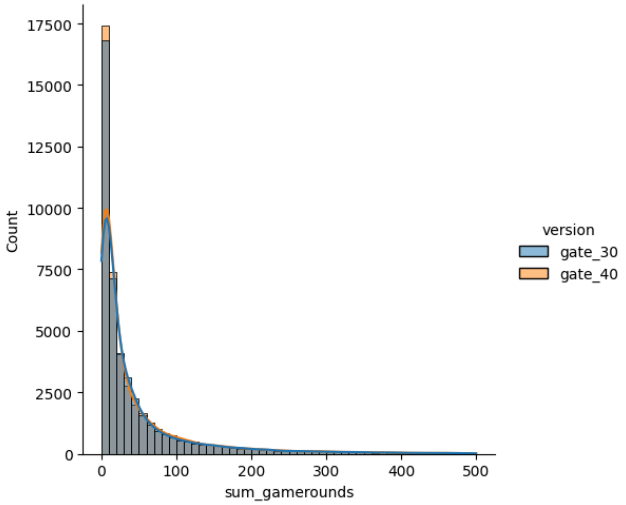
Cookie Cats is an ideal dataset for function validation because it includes both binary and continuous metrics, extremely skewed distributions, a large sample size (90k users), and real A/B test behavior observed in prior analyses.

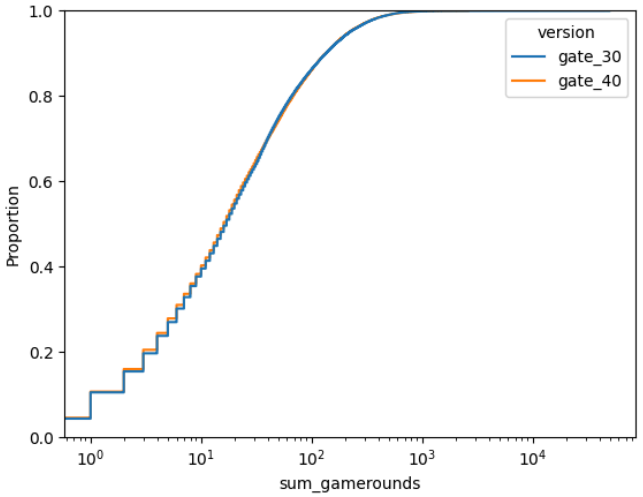
1. Dataset Overview

The dataset contains gameplay data from 90,189 users, including:

Column	Description
userid	unique player ID
version	gate_30 (control) or gate_40 (treatment)
sum_gamerounds	total rounds played in 14 days
retention_1	returned next day (binary)
retention_7	returned within 7 days (binary)

There are no missing values or duplicates.





sum_gamerounds is heavily right-skewed, but with ~45k users per group, the Central Limit Theorem ensures that parametric tests are valid for comparing means.

Initial EDA shows that group means and distributions are remarkably similar.

2. Internal Validity

Sample Ratio Mismatch (SRM)

- A chi-square test showed a small but statistically significant imbalance ($p \approx 0.0086$).
- The absolute deviation was <1%, which is negligible at this scale.
- Given that this dataset is curated for teaching, this likely reflects prior filtering, not a true randomization failure.

A/A Test

- The control group (gate_30) was randomly split into two fake arms (A1 & A2).
- The A/A test showed no meaningful differences in engagement or retention.
- Zero variance in one split led to nan p-values for some metrics, but overall behavior between A1 and A2 remained aligned — indicating no major measurement issues.

3. External Validity

Simpson's Paradox Check

- I created random, pre-treatment segments using: df["segment"] = df["userid"] % 3
- This checks whether treatment effects reverse in different user slices.

Results:

- **Retention metrics** were consistent across all segments
- No reversal patterns for retention

- **sum_gamerounds** reversed in two segments
    - Expected due to its extreme variance and heavy-tailed distribution
    - Common in gameplay data where a small group of “heavy players” dominate totals

Overall, retention metrics show robust behavior; engagement metrics require caution.

4. Hypothesis Testing

Metric	P-value	Effect Size	Test Type	Significant	Effect Size Label
Ratio game rounds (sum_gamerounds / 7)	0.376	-0.083	delta	False	Raw difference
Retention 1	0.074	0.012	proportion	False	Very small
Retention 7	0.002	0.021	proportion	True	Very small
Sum game rounds	0.378	0.006	mean	False	Very small

5. Power Analysis

Using conventional thresholds:

- Proportion effects observed here (<2% difference) require >300k users per group to detect reliably.
- Our sample (~45k per group) is significantly underpowered for detecting such tiny effects.

Thus: The experiment is statistically underpowered for the effect sizes observed.

6. Conclusion & Recommendation

- Only retention_7 showed statistical significance
- But the practical effect is negligible
- sum_gamerounds and ratio metrics show no meaningful impact
- Simpson’s paradox suggests instability in high-variance engagement metrics
- Power analysis confirms the experiment is too small to detect small behavioral changes

Recommendation:

- Do not move the gate to level 40 based on this experiment.
- A larger-scale re-test is necessary if the product team still wants to evaluate this design change.