Mobile App A/B Testing

Cookie Cats "gate" version comparison

Context

"Cookie Cats is a hugely popular mobile puzzle game developed by Tactile Entertainment. It's a classic "connect three" style puzzle game where the player must connect tiles of the same color in order to clear the board and win the level."
   - Rasmus Bååth

Cookie Cats places "gates" after certain levels throughout the game that forces players to wait a certain period of time before advancing, or pay to move on early. This serves two purposes:

1. Increase revenue through in-game purchases.
2. Increase player satisfaction through forced breaks.

Research Question

"A/B testing is the most powerful way to turn clicks into customers."
- Dan Siroker

Currently, the developers of Cookie Cats have placed the game's first gate at level 30. They would like to know the potential impact of moving the first gate to level 40. A/B testing is a powerful and flexible tool that would be great for this use case to compare the two groups. For this analysis, I will be using both continuous and categorical A/B testing methods to analyze the impact of the gate change.

Data Preparation

Upon downloading the mobile app, users were randomly assigned to receive their first gate either after level 30 or level 40. Three variables of interest were gathered over the course of their first week:

1. Total game rounds played.
2. 1 day retention status.
3. 7 day retention status.

The table below reveals that there are users who never played a game after downloading the game, who never made it to either of the gate options, and users who played a suspiciously large number of games. These users will be removed from the analysis to ensure the reliability of the results.

Rounds Played Summary Statistics

| Min	1st Qu.	Median	Mean	2nd Qu.	Max
0	5	16	51.87	51	49854

Continuous Analysis

Data visualizations:

1. The distribution of users assigned to each gate version upon downloading the mobile app.
2. A truncated view of the distribution of the user counts by total games played within the first week of downloading the mobile app.

Insights from the graphs:

- There is an even distribution of users assigned to each gate version.
- The continuous response variable, total rounds played, is right skewed.
  - This would indicate that we should use a non-parametric statistical test to compare the two groups.

User Counts per Gate Version



User Counts per Total Rounds Played
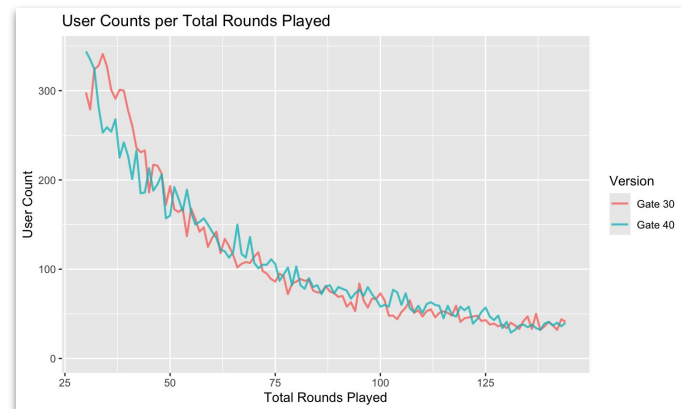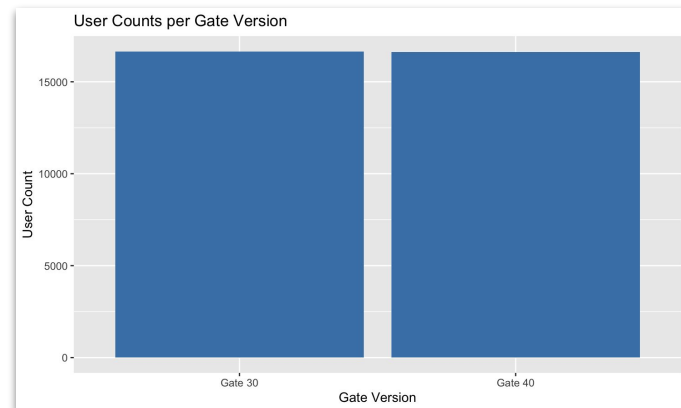
Continuous Analysis

Mann-Whitney U test:

- Hypotheses
    - Null hypothesis: Shifting the gate from level 30 to level 40 does not impact the number of rounds a user plays during their first week.
    - Alternative hypothesis: Shifting the gate from level 30 to level 40 does impact the number of rounds a user plays during their first week.
- Results
    - p-value = 1×10^{-4}
    - Reject the null hypothesis: shifting the gate from level 30 to level 40 does impact the number of rounds a user plays during their first week ($\alpha = 0.05$).
- Follow-up / Interpretation
    - Now that we now there is a statistical difference between the two gate versions, we need to know how they differ.
    - In order to do this, I will bootstrap the mean for each group 10,000 times with replacement and plot the bootstrapped means alongside a difference distribution (Gate 40 - Gate 30).

Density Plots for Mean Rounds Played
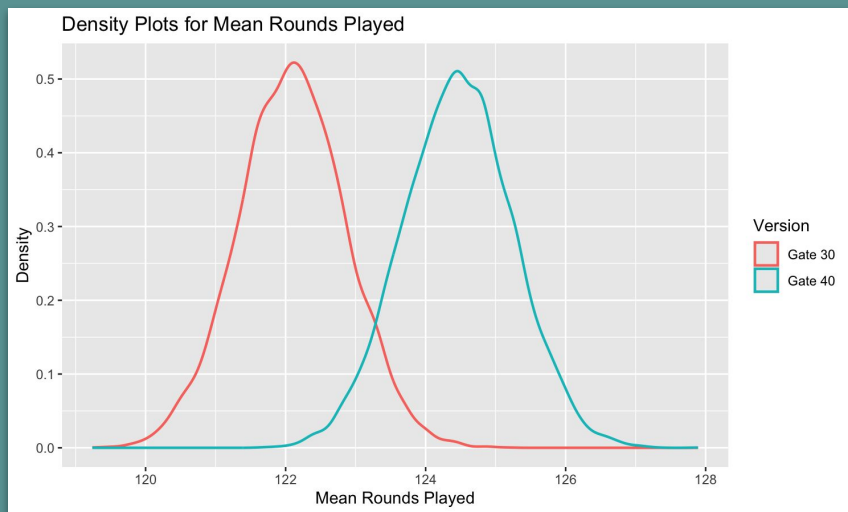
The density distribution of bootstrapped means.

Histogram of Mean Differences (Gate 40 - Gate 30)
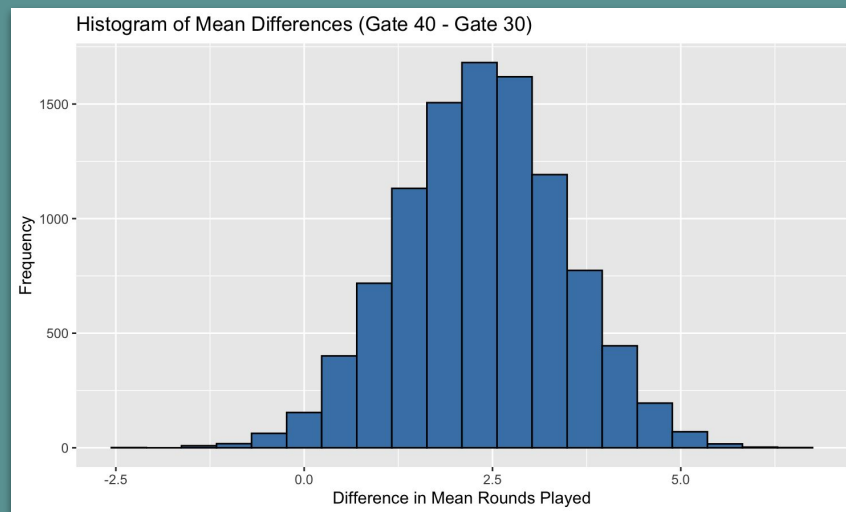
The difference distribution of bootstrapped means.

Average difference: **+2.4 games**

Categorical Analysis

The total number of rounds played is not the only variable of interest to the developers; they are also interested in the retention rates of these players. To analyze this, I will conduct two Chi-square tests to examine the retention rate counts for both gate versions: one for the 1-day retention rate and another for the 7-day retention rate.

First, we must check the assumption of a Chi-squared test that there are at least 5 observations in each category. The tables below demonstrate that the data satisfy this.

1 Day Retention User Counts

	True	False
Gate 30	20,034	24,666
Gate 40	20,119	25,370

7 Day Retention User Counts

	True	False
Gate 30	8,502	36,198
Gate 40	8,279	37,210

Categorical Analysis

Chi-squared test of independence:

- Hypotheses
  - Null hypothesis: There is no relationship between retention rates and gate version.
  - Alternative hypothesis: There is a relationship between retention rates and gate version.
- Results
  - 1 day retention p-value: 0.076
  - 7 day retention p-value: 0.002
  - Fail to reject the null hypothesis for 1 day retention rates, reject null hypothesis for 7 day retention rates ($\alpha = 0.05$).
- Follow-up
  - Because of the simplicity of the categorical analysis, we only need to compare the retention rates between the two gate versions to draw conclusions; no need for bootstrapping or a post-hoc statistical test.
  - 7 day retention rates:
    - Gate 30: 19.0%
    - Gate 40: 18.2%

Conclusion

- Increasing the first gate users encounter to level 40 shows a statistically significant increase in the number of rounds played in their first week of playing.
- Conversely, Increasing the first gate to level 40 shows a statistically significant decrease in 7 day retention rates among new users.
  - However, this shift is small (0.8%), and might not bear much practical significance.
- Recommended course of action: *Increase the first gate to level 40.*