Title: A/B Testing for Mobile Games - Cookie Cats

Goal

The goal of this A/B test is to assess the impact of moving the first gate in the Cookie Cats mobile game from level 30 to level 40. Specifically, the goal is to determine how this change affects player engagement, retention rates (both after 1 day and 7 days), and the number of game rounds played within the first 14 days after installation.

Experiment Design

Contents

The dataset consists of data from 90,189 players who installed the game during the A/B test period.

Task Variables

The variables provided in the dataset include:

user-id: A unique number that identifies each player.
version: Whether the player was put in the control group (gate_30 - a gate at level 30) or the group with the moved gate (gate_40 - a gate at level 40).
sum_gamerounds: the number of game rounds played by the player during the first 14 days after installation.
retention_1: Did the player come back and play **1 day** after installing?
retention_7: Did the player come back and play **7 days** after installing?

When a player installed the game, he or she was randomly assigned to either.

Metrics Choice

I used BigQuery to run queries and generate key metrics required for this analysis.

The variables are

- game_version;
- no_of_users,

- no_of_retention_1;
- retention_1_rate;
- no_of_retention_7;
- retention_7_rate;
- Mean_of_gamerounds,
- stddev_of_gamerounds,
- min_value
- max_value

Check this spreadsheet or the image below.

Note: gate_30 is the control group; and
gate_40 is the experiment group

game_version	no_of_users	no_of_retention_1	retention_1_rate	no_of_retention_7	retention_7_rate	mean_of_gamerounds	stddev_of_gamerounds	min_value	max_value
gate_30	44700	20034	0.4482	8502	0.1902	52.4563	256.7164	0	49854
gate_40	45489	20119	0.4423	8279	0.1820	51.2988	103.2944	0	2640

In the results shown above, the retention_1 rate and retention_7 rate for the gate_30 group are higher than that of the gate_40 group. Additionally, the standard deviation of game rounds and the mean of game rounds are also higher than that of the gate_40 group, further indicating that the gate_30 group performed better overall compared to the gate_40 group.

However, this is not sufficient enough to launch the features. Further analysis needs to be conducted on the evaluation metrics to guide our actions.

Hypothesis

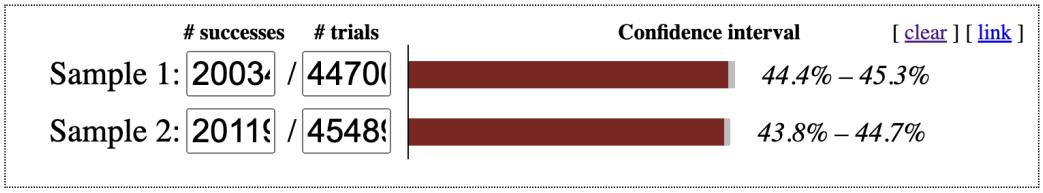
- Null hypothesis (H_0): The retention rate is the same for both groups (no difference). I.e
$H_0 : p1=p2$

- Alternative hypothesis (H_a): The retention rates are different between the two groups. i.e
$Ha : p1 \neq p2$

For my next step, I used Evan Millers's Chi-Squared Test and the 2 Sample T-Test calculator to further carry out my analysis.

Evan Millers's Chi-Squared Test.

Retention_1:



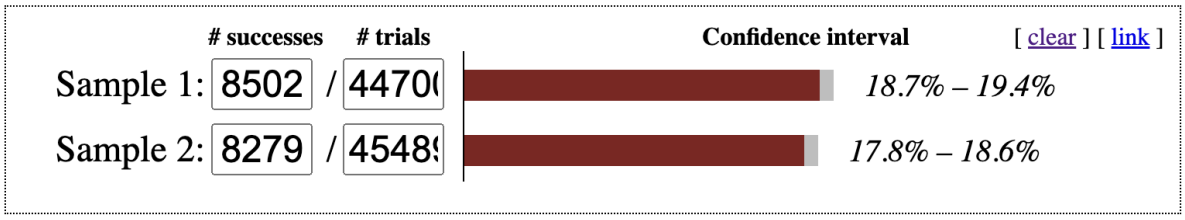
Verdict:

No significant difference

(p = 0.074)

Confidence level: 95%

Retention_7:



Verdict:

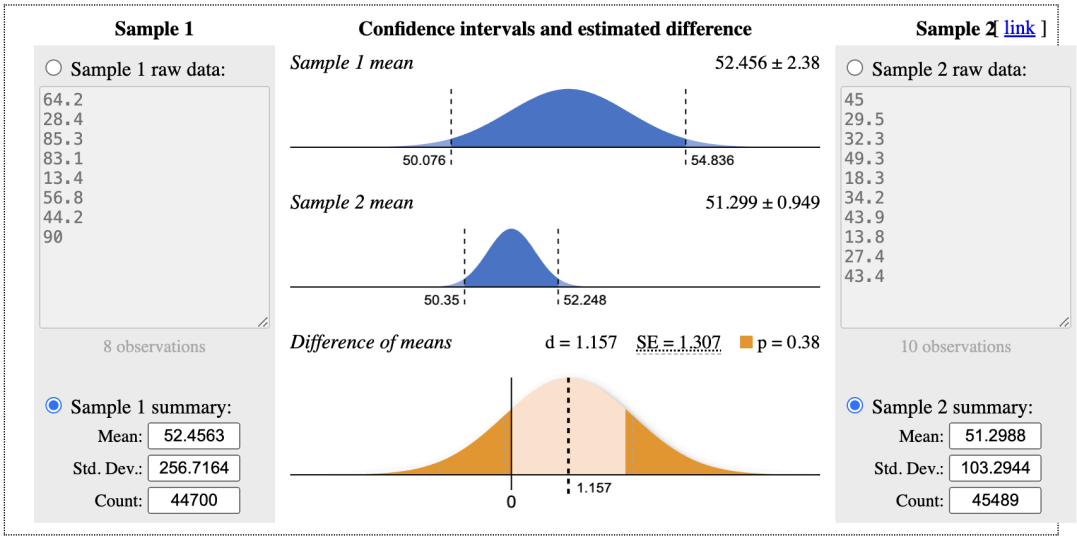
Sample 1 is more successful

(p = 0.00155)

Confidence level: 95%

Evan Millers's Chi-Squared Test (95% CI)		
	p-value	Statistically significant?
Retention_1	0.074	No
Retention_7	0.00155	Yes

Evan Millers's 2 Sample T-Test



Verdict: No significant difference

Hypothesis: d = 0 d ≤ 0 d ≥ 0

Confidence: 95%

2 Sample T-Test (95% CI)	
p-value	0.38
Statistically significant?	No

1. Description of Statistical Tests:

- **2-Sample T-Test**: A 2-sample t-test was conducted to compare the mean number of game rounds played between Gate 30 and Gate 40.
- **Chi-Squared Tests**: Two separate Chi-Squared tests were conducted to assess the differences in retention rates:
    - **Retention after 1 day**: The chi-squared test yielded a p-value of 0.074.
    - **Retention after 7 days**: The chi-squared test yielded a p-value of 0.0016.

2. Estimated Treatment Effect:

- **Difference in Means (Game Rounds)**: The estimated difference in mean game rounds between Gate 30 (Sample 1) and Gate 40 (Sample 2) is 1.157, with Gate 30 players playing more rounds on average.

3. Confidence Intervals:

- **Gate 30 (Game Rounds)**: 52.456 ± 2.38 (50.076 to 54.836).
- **Gate 40 (Game Rounds)**: 1.299 ± 0.949 (0.350 to 2.248).
- These confidence intervals indicate that Gate 30 has a higher mean game round than Gate 40.

4. P-Values:

- **2-Sample T-Test**: The p-value is 0.38, indicating no statistically significant difference in the mean number of game rounds between Gate 30 and Gate 40.
- **Chi-Squared Test (Retention after 1 day)**: The p-value is 0.074, which is above the standard alpha level of 0.05. While this suggests some difference in day 1 retention between Gate 30 and Gate 40, it is not statistically significant.
- **Chi-Squared Test (Retention after 7 days)**: The p-value is 0.0016, which is well below 0.05, indicating a statistically significant difference in retention rates after 7 days, with Gate 30 showing better long-term retention.

Summary Statement

- **Retention after 1 day**: The p-value of 0.074 suggests that the difference between Gate 30 and Gate 40 is marginally insignificant. While Gate 30 has a slightly higher retention rate (44.82% vs. 44.23%), this result is not statistically significant.
- **Retention after 7 days**: With a highly significant p-value of 0.0016, Gate 30 clearly outperforms Gate 40 in long-term retention (19.02% vs. 18.20%). This is a crucial metric, as long-term retention typically indicates stronger player engagement.
- **Game Rounds Played**: Although Gate 30 has a higher mean number of game rounds (52.46 vs. 51.30), the difference is not statistically significant ($p = 0.38$). This suggests that the change in the gate position does not drastically affect overall player engagement measured by game rounds.

Recommendation

Given that the retention rates after 7 days show a statistically significant improvement for Gate 30, this should be the preferred option, even though the retention after 1 day and the number of game rounds played are not significantly different. Long-term player engagement is typically more valuable, making Gate 30 a better choice for maximizing retention.

Action: **Leave Gate 30 as the first gate** as it shows stronger long-term player retention, which is likely to yield more sustained engagement over time.

Appendix

SQL Query

```
WITH

 selected_data AS (

 SELECT

   version AS game_version,

   COUNT (DISTINCT user_id) AS no_of_users,

   SUM (CASE

       WHEN retention_1 = TRUE THEN 1

       ELSE 0

   END

     ) AS no_of_retention_1,

   SUM (CASE

       WHEN retention_2 = TRUE THEN 1

       ELSE 0

   END

     ) AS no_of_retention_7,

   AVG(sum_gamerounds) AS mean_of_gamerounds,

   STDDEV(sum_gamerounds) AS stddev_of_gamerounds,

   MIN(sum_gamerounds) AS min_value,

   MAX(sum_gamerounds) AS max_value
```

```
 FROM
   `tc-da-1.turing_data_analytics.cookie_cats`
 GROUP BY
   game_version),
 percentiles_data AS (
 SELECT
   version AS game_version,
   PERCENTILE_CONT(sum_gamerounds, 0.50) OVER (PARTITION BY version) AS
median_of_gamerounds,
   PERCENTILE_CONT(sum_gamerounds, 0.99) OVER (PARTITION BY version) AS
p99_of_gamerounds
 FROM
   `tc-da-1.turing_data_analytics.cookie_cats`)
SELECT
 selected_data.game_version,
 no_of_users,
 no_of_retention_1,
 no_of_retention_1/no_of_users AS retention_1_rate,
 no_of_retention_7,
 no_of_retention_7/no_of_users AS retention_7_rate,
 mean_of_gamerounds,
 stddev_of_gamerounds,
 min_value,
```

```
 median_of_gamerounds,
 p99_of_gamerounds,
 max_value
FROM
 selected_data
INNER JOIN
 percentiles_data
ON
 selected_data.game_version = percentiles_data.game_version
GROUP BY
 selected_data.game_version,
 no_of_users,
 no_of_retention_1,
 no_of_retention_7,
 mean_of_gamerounds,
 min_value,
 stddev_of_gamerounds,
 median_of_gamerounds,
 p99_of_gamerounds,
 max_value
```

Problem source: Kaggle