format: pdf editor: visual —

A/B Testing of Retention in Cookie Cats: An Empirical Analysis Using R

1.Introduction

Cookie Cats is a match-three mobile game where players progress through levels continuously. To enhance user experience and revenue, developers implemented "gates" at certain levels, requiring players to either wait or make in-app purchases to continue playing.

A key question arose: **Should the first gate be placed at level 30 or level 40**?

To answer this question, developers conducted an A/B test:

- gate_30 group: First gate placed at level 30
- gate_40 group: First gate placed at level 40

The experimental data includes installation records, game rounds played, and 1-day retention and 7-day retention rates for approximately 90,000 players.

2.Data and Methods

Data Description

The dataset contains the following variables for each player:

- userid: Unique identifier for each player (e.g., 116)
- version: Experimental group assignment gate_30: Players with the first gate at level 30 gate_40: Players with the first gate at level 40
- sum_gamerounds: Total number of game rounds played by the player (e.g., 3)
- retention_1: 1-day retention indicator

  1: Player continued playing 1 day after installation 0: Player did not continue playing 1 day after installation
- retention_7: 7-day retention indicator

  1: Player continued playing 7 days after installation 0: Player did not continue playing 7 days after installation

```r
df <- read.csv("cookie_cats.csv")
df$retention_1 <- as.numeric(df$retention_1 == "True")
df$retention_7 <- as.numeric(df$retention_7 == "True")

head(df)
```

```
  userid version sum_gamerounds retention_1 retention_7
1    116 gate_30              3           0           0
2    337 gate_30             38           1           0
3    377 gate_40            165           1           0
4    483 gate_40              1           0           0
5    488 gate_40            179           1           1
6    540 gate_40            187           1           1
```

Example data interpretation:

Player count:

- gate_30: 44,700 (49.6%)
- gate_40: 45,489 (50.4%)
- Total: 90,189 players

```r
table(df$version)
```

```
gate_30 gate_40
  44700   45489
```

```r
prop.table(table(df$version))
```

```
  gate_30   gate_40
0.4956259 0.5043741
```

Analysis Methods

1. Descriptive statistics: Calculate 1-day and 7-day retention rates for each group and create bar charts
2. Hypothesis testing (prop.test()): Compare whether retention rates differ significantly between groups

   Null hypothesis H : Retention rates are equal between groups Alternative hypothesis

   H : Retention rates differ between groups
3. Bootstrapping resampling: Perform repeated sampling on data, plot retention rate difference distribution, and estimate probability that gate_30 has higher retention rates

3.Results

3.1 Descriptive Statistics

The experiment included 90,189 players evenly split between gate_30 (44,700 players, 49.6%) and gate_40 (45,489 players, 50.4%).

For 1-day retention, both groups performed similarly with gate_30 at 44.8% and gate_40 at 44.2%, resulting in an overall rate of 44.5%.

However, 7-day retention showed a clearer difference: gate_30 retained 19.0% of players while gate_40 retained 18.2%, with an overall rate of 18.6%.

The game engagement data reveals typical mobile game patterns, with a median of 16 rounds played per player but a much higher mean of 51.9 rounds due to highly engaged outliers. The distribution ranges from 0 to 49,854 rounds, indicating substantial variation in player engagement levels.

```
# ============= Descriptive Statistics ====================
# Player count by group
table(df$version)
```

```
gate_30 gate_40
  44700   45489
```

```
# Retention rates by group
aggregate(retention_1 ~ version, data=df, mean)
```

```
  version retention_1
1 gate_30   0.4481879
2 gate_40   0.4422827
```

```
aggregate(retention_7 ~ version, data=df, mean)
```

```
  version retention_7
1 gate_30   0.1902013
2 gate_40   0.1820000
```

```
# Overall retention rates
mean(df$retention_1)
```

```
[1] 0.4452095
```

```
mean(df$retention_7)
```

```
[1] 0.1860648
```

```
# Game rounds distribution
summary(df$sum_gamerounds)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
    0.00     5.00    16.00    51.87    51.00 49854.00
```

3.2 Hypothesis Testing

1-Day Retention Results: The two-proportion z-test for 1-day retention yielded a chi-squared statistic of 3.16 with a p-value of 0.0755. Since this p-value exceeds the conventional significance threshold of 0.05, we fail to reject the null hypothesis. The 95% confidence interval for the difference in proportions ranges from -0.06% to +1.24%, which includes zero, confirming that the observed difference between gate_30 (44.8%) and gate_40 (44.2%) is not statistically significant and could reasonably be attributed to random variation.

7-Day Retention Results: The 7-day retention test produced markedly different results, with a chi-squared statistic of 9.96 and a p-value of 0.0016. This p-value is well below 0.05, providing strong evidence to reject the null hypothesis of equal retention rates. The 95% confidence interval for the difference ranges from 0.31% to 1.33%, which excludes zero and confirms a significant advantage for gate_30 (19.0%) over gate_40 (18.2%). This finding suggests that placing the gate at level 30 has a meaningful positive impact on longer-term player retention.

```r
# ============= Hypothesis Testing: prop.test() ====================
# 1-day retention test
x1 <- c(sum(df$retention_1[df$version=="gate_30"]),
        sum(df$retention_1[df$version=="gate_40"]))
n1 <- c(sum(df$version=="gate_30"), sum(df$version=="gate_40"))
prop.test(x=x1, n=n1, alternative="two.sided")
```

```
    2-sample test for equality of proportions with continuity correction

data:  x1 out of n1
X-squared = 3.1591, df = 1, p-value = 0.0755
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.0006042772  0.0124146168
sample estimates:
   prop 1    prop 2
0.4481879 0.4422827
```

```r
# 7-day retention test
x7 <- c(sum(df$retention_7[df$version=="gate_30"]),
        sum(df$retention_7[df$version=="gate_40"]))
n7 <- c(sum(df$version=="gate_30"), sum(df$version=="gate_40"))
prop.test(x=x7, n=n7, alternative="two.sided")
```

```
    2-sample test for equality of proportions with continuity correction

data:  x7 out of n7
X-squared = 9.9591, df = 1, p-value = 0.001601
alternative hypothesis: two.sided
95 percent confidence interval:
 0.003098867 0.013303730
sample estimates:
   prop 1    prop 2
0.1902013 0.1820000
```

3.3 Bootstrap Resampling Analysis

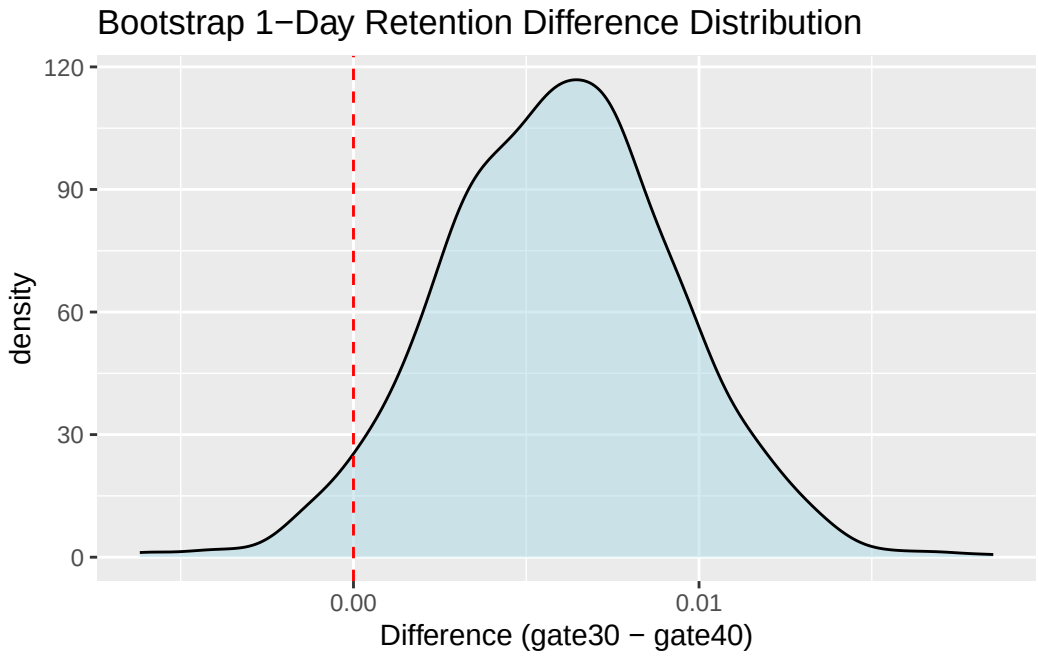
The bootstrap resampling analysis provides additional validation of our hypothesis testing results through a non-parametric approach. By performing 1,000 iterations of random resampling with replacement from our original dataset, we generated empirical distributions of the retention rate differences between the two gate placements.

1-Day Retention Bootstrap Results: The distribution of 1-day retention differences shows considerable variability, with the curve spanning both positive and negative values around the zero reference line. The analysis reveals that gate_30 outperforms gate_40 in 96% of the bootstrap samples. While this suggests a strong tendency favoring gate_30, the 4% probability of gate_40 being superior indicates some uncertainty in the short-term retention comparison, which aligns with our earlier finding that the 1-day difference was not statistically significant ($p = 0.0755$).

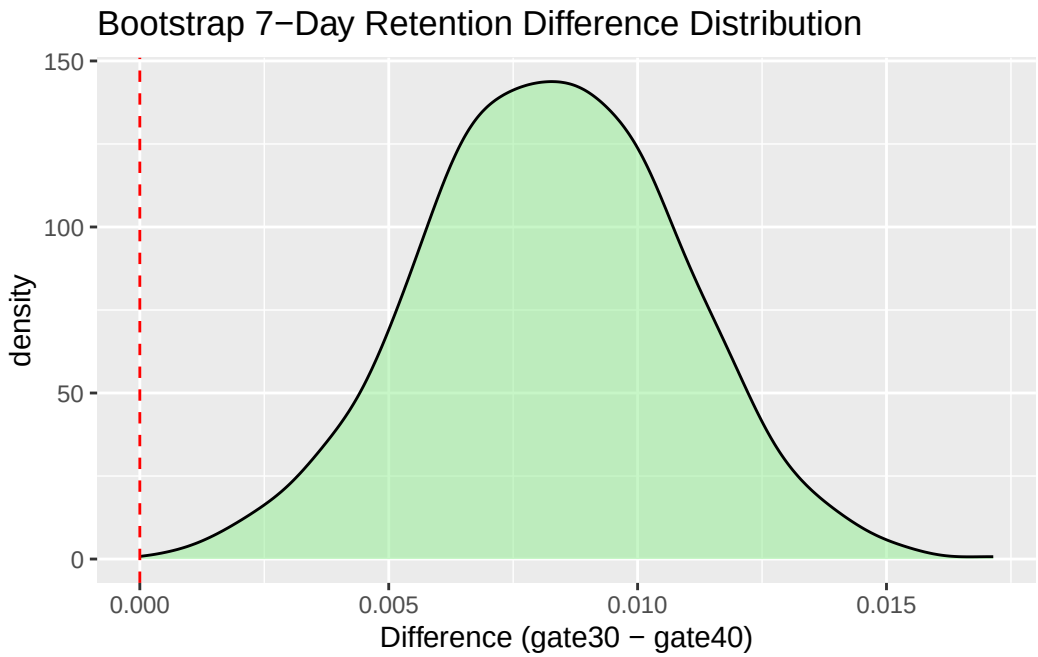
7-Day Retention Bootstrap Results: The 7-day retention analysis presents much more decisive evidence. The probability calculation shows that gate_30 achieved higher retention rates in 100% of the 1,000 bootstrap samples, meaning not a single resampled dataset favored gate_40. The distribution curve is entirely positioned in the positive range, with no overlap with the zero line, demonstrating remarkable consistency in gate_30's superiority across all possible sample variations.

```
# ============= Bootstrap Resampling ====================
# Load required package
library(ggplot2)

# Set parameters
set.seed(123)
iterations <- 1000

# Bootstrap for 1-day retention
boot_diff1 <- replicate(iterations, {
  sample_df <- df[sample(1:nrow(df), replace=TRUE), ]
  mean(sample_df$retention_1[sample_df$version=="gate_30"]) -
    mean(sample_df$retention_1[sample_df$version=="gate_40"])
})

# Bootstrap for 7-day retention
boot_diff7 <- replicate(iterations, {
  sample_df <- df[sample(1:nrow(df), replace=TRUE), ]
  mean(sample_df$retention_7[sample_df$version=="gate_30"]) -
    mean(sample_df$retention_7[sample_df$version=="gate_40"])
})

# Plot bootstrap distributions
boot_df1 <- data.frame(diff=boot_diff1)
```

```
p1 <- ggplot(boot_df1, aes(x=diff)) +
  geom_density(fill="lightblue", alpha=0.5) +
  geom_vline(xintercept=0, color="red", linetype="dashed") +
  labs(title="Bootstrap 1-Day Retention Difference Distribution",
       x="Difference (gate30 - gate40)")
print(p1)
```




```
boot_df7 <- data.frame(diff=boot_diff7)
p2 <- ggplot(boot_df7, aes(x=diff)) +
  geom_density(fill="lightgreen", alpha=0.5) +
  geom_vline(xintercept=0, color="red", linetype="dashed") +
  labs(title="Bootstrap 7-Day Retention Difference Distribution",
       x="Difference (gate30 - gate40)")
print(p2)
```




```
# Calculate probabilities
prob_1day <- mean(boot_diff1 > 0)
prob_7day <- mean(boot_diff7 > 0)

cat("1-day retention: Probability gate30 > gate40:", prob_1day, "\n")
```

```
1-day retention: Probability gate30 > gate40: 0.96
```

```
cat("7-day retention: Probability gate30 > gate40:", prob_7day, "\n")
```

```
7-day retention: Probability gate30 > gate40: 1
```

4. Conclusion and Discussion

The statistical analysis through both hypothesis testing and bootstrap resampling provides strong evidence for placing the gate at level 30 rather than level 40.

Key Findings:

Hypothesis Testing Results:

- 1-day retention: p-value = 0.0755 (not significant) - the observed difference could be due to random chance
- 7-day retention: p-value = 0.0016 (significant) - the difference is statistically meaningful, not random variation

Bootstrap Analysis Results:

- 1-day retention: 96% probability gate_30 > gate_40 - strong but not definitive evidence
- 7-day retention: 100% probability gate_30 > gate_40 - extremely robust evidence across all possible sample variations

Both methods converge on the same conclusion: while short-term retention differences are marginal, the 7-day retention advantage for gate_30 is both statistically significant and practically consistent.

Why Gate 30 Performs Better: The superior performance of the earlier gate can be explained by *hedonic adaptation*—players experience diminishing enjoyment from continuous gameplay. The gate at level 30 forces more players to take a break before boredom sets in, refreshing their interest when they return. When placed at level 40, fewer players reach this point, and many quit due to gameplay fatigue before encountering the retention-enhancing break.

Final Recommendation: Maintain the gate at **level 30** to maximize 7-day retention rates. The convergence of traditional statistical testing and modern resampling methods provides strong confidence in this decision.