BACHELOR'S DEGREE THESIS

Degree in Statistics

Title: Assessing and reducing bias for a Bayesian AB test

Author: Auba Vásquez Carreras

Director: Joan Bas Serrano

Advisor: Víctor Peña Pizarro

Department: Department of Statistics and Operations Research

Academic year: 2025

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Abstract

Working with data from Free To Play videogames, we usually have heavy-tailed distributions that can be quite complex to study. When performing AB tests to evaluate certain changes in the games, the different users' spending behaviour might lead to a pretest bias that can affect the results. Our aim is to find a way to analyze the tests that considers the information from the users before entering the test but also takes into account the pretest bias. We will study the current method used to analyze AB tests at the company and identify some of its problems, such as the use of the probability to beat the control group as a metric. We will also explore some variance reduction techniques, such as CUPED, to get better results focusing on the ARPU and Conversion.

Key Words

Pretest bias, AB test, ARPU, Conversion, CUPED, Variance, Bayesian statistics.

Resum

Quan treballem amb dades de jocs gratuïts, generalment trobem distribucions amb cues molt pesades que poden ser complexes d'estudiar. En dur a terme tests AB per avaluar canvis en els jocs, els diferents comportaments dels usuaris quant a les despeses poden generar un biaix abans del test que pot afectar els resultats. El nostre objectiu és analitzar els tests considerant les dades prèvies al test però també tenint en compte el biaix. Estudiarem el mètode actual utilitzat a l'empresa per analitzar tests AB i n'identificarem alguns problemes, com l'ús de la mètrica probabilitat de guanyar el grup de control. També examinarem alguns mètodes per reduir la variància, com CUPED, per obtenir millors resultats en l'anàlisi, centrant-nos en les mètriques *ARPU* i Conversió.

Paraules clau

Biaix abans del test, Test AB, ARPU, Conversió, CUPED, Variància, Estadística Bayesiana.

Mathematics Subject Classification (AMS)

62F15 Bayesian inference
62P30 Applications in engineering and industry

Acknowledgements

This project marks the end of a stage in my life. These last years have been filled with ups and downs and I could not have made it here alone. So I want to thank everyone who has been a part of this journey.

> To Joan and Víctor, for their involvement in this project and constant support these last months.
>
> To the Analytics department, for making me feel part of the team since the first day and for trusting me with this project.
>
> To my friends, for always being by my side. Especially my friends from university, for being much more than classmates. And my friends from Menorca, for making me feel at home wherever we are.
>
> Finally, to my parents, for believing in me even when I don't.

To all of you, thank you.

Index

Introduction

An AB test is a controlled experiment in which we try to see if a certain change in the game significantly affects the users' spending behaviour. We take a group of users and divide them into two groups: the test group that will use a new version with the change we want to study and the control group that serves as the baseline. The problem with Free To Play videogames is that there is usually a small percentage of paying users and, within this group, an even smaller percentage of users that contribute to a large part of the total revenue. Thus, we have heavy-tailed distributions that are quite complex to study. Moreover, with bayesian AB testing we also want to consider pretest data, but these high-spending users tend to make this quite complicated, since having them in one group or the other can lead to a significant pretest bias that might provide some misleading results.

In this dissertation we will study the effects of the pretest bias on AB tests, as well as some methods that might reduce its effects. For this, first we will perform an exploratory data analysis to get a clear picture of the distributions and the spending behaviour of the users. Then we will reproduce the current analysis method to simulate AB tests and study the metrics obtained, as well as identify some key problems of this processing. After that, we will suggest new analysis methods and study them to see if the results obtained improve the current ones.

This project is carried out within the framework of an internship, so the data used for this project corresponds to user revenue information for mobile video games of the company. We will use data from three different games and focus on In App Purchases. Our main goal is to obtain a better way to analyze AB tests in the company that accounts for the effects of the pretest bias.

Glossary

ARPU: Average Revenue Per User. Considering a group of users, the average of all their revenues during a certain period of time.

Conversion: For a user, boolean metric that takes the value 1 if the user has done any purchase in the time period considered and 0 otherwise. For a group of users, the average of all their conversions.

F2P: Free to Play. Games that can be downloaded and played without an initial purchase, but often offer IAP

IAP: In App Purchases

KPI: Key Performance Indicator

NGU: New Game User

Methods

Current approach

The current method used to analyze test results only takes into account the data collected during the test, ignoring the previous information. So, it takes the AB test data and, for each group (control or test) generates a sample of 2 000 values of the metric. Averaging these samples, we obtain the value for each group and, with these, we can then calculate the metrics we need for our report: uplift, credible interval for the uplift, probability to beat the control group, uplift if the group is better and uplift if it is worse. The samples are generated following a different distribution depending on the metric we are working with.

If we are working with the **Conversion**, since it is a proportion, we use a Beta distribution as a prior and Binomial for the likelihood, obtaining a Beta-Binomial as a posterior distribution. Since it is a conjugate model, it can be simulated directly.

However, when working with the **ARPU**, since it is a numerical KPI, the distribution used is a NormalGamma-Normal. With the prior distribution being a Normal-Gamma and the likelihood following a Normal distribution.

Mixture model

Furthermore, we will try a new model for the ARPU, a mixture model. In this case, we will be modeling one specific distribution at a time, which can represent a single point in time for a single group (a specific combination in {pre, post} x {control, test}). This model will be adjusted using the JAGS package in R.

Given that we are modeling a process where many observations are zeros (non-payers), we will use a zero-inflated model that has two components.

1. Binary component:
   Each observation y_i has an associated latent indicator z_i defined as:

   $$z_i \sim Bernoulli(1 - p)$$

   where p is the probability of a structural zero (a non-spender). With probability p the observation is zero, and with probability $1 - p$ the observation is non-zero.

2. Continuous component:
   For non-zero observations (when z_i=1), we model a log-normal distribution. This means that the logarithm of spending follows a normal distribution:

$$log(y_i) \sim N(\mu, \sigma^2)$$

   which implies that y_i is distributed as a log-normal:

$$y_i \sim LogNormal(\mu, \sigma^2)$$

We use non-informative priors for the parameters to reflect our lack of strong knowledge and to allow the data to drive the inference.

- Zero Probability

$$p \sim Beta(1, 1)$$

  This is equivalent to a uniform prior on the interval [0,1], which is diffuse (non-informative) and does not favor any particular value

- Mean of Log-Revenue

$$\mu \sim N(5, 10^2)$$

  A large variance (10^2) expresses considerable uncertainty about the mean value.

- Variance for Log-Revenue

$$\sigma^2 \sim Inverse - Gamma(2, 0.1)$$

  It should be taken into account that for clarity we are parametrizing the distribution in terms of variance (σ) instead of precision ($\tau=1/\sigma^2$), which is commonly used in Bayesian statistics.

1. Exploratory Data Analysis

Our first step will be a general exploratory analysis of the data we will be working with to get a general idea of its structure and behaviour. We will start analyzing the data for one specific game (Game 1) for a two month period. With a histogram and a boxplot we can see that there is a considerable amount of users with very low (or null) revenues, which is very common for F2P games.

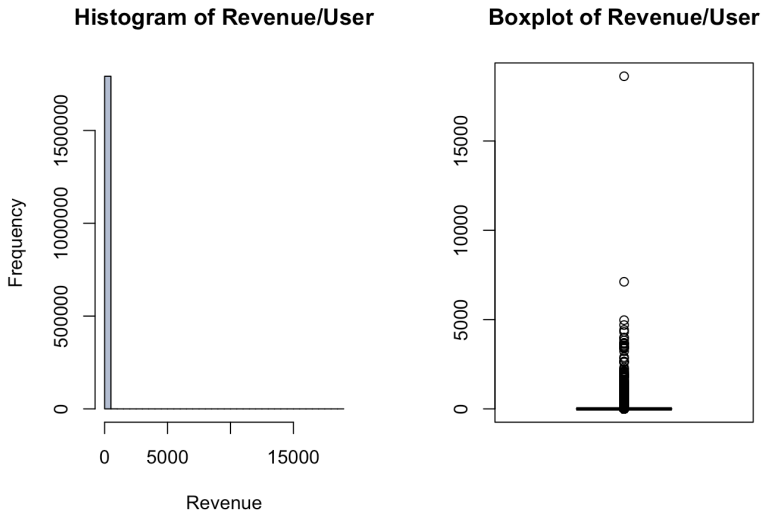

Figure 1: Histogram and boxplot of the data for a two month period for Game 1.

To get a deeper understanding of the data, we separate the users into four different buckets depending on their revenue:

- No revenue
- Up until 10 USD
- More than 10 USD, up to 100 USD
- More than 100 USD

We can clearly see that the percentage of non-paying users is higher than 95% for all games and, in all cases, the smallest group is the one formed by users that have spent more than 100USD in the time period considered.

	% of users per bucket			% of IAP revenue		
	G1	**G2**	**G3**	**G1**	**G2**	**G3**
0	96.68	97.53	95.33	0	0	0
(0, 10]	1.85	1.51	2.33	5.52	8.99	5.19
(10, 100]	1.22	0.84	1.91	34.72	41.46	34.30
(100, +∞]	0.24	0.11	0.43	59.76	49.54	60.50

Table 1: Percentage of users per bucket and contribution to the IAP revenue, divided by spending buckets

1.1 Sampling from the data

For the next step of the Exploratory Data Analysis, instead of looking at all the users, we will take samples and analyze their metrics before and after a set date, simulating an AB test (or AA test, since all the data comes from the same population). This will also help us get a deeper insight into how the pretest bias affects our analysis. For each user, we have their IAP revenue from one month (equivalent to the pre-test period) and their revenue from the next month (the test period, that we will refer to as “post”), adding up to around 460 000 observations.

First of all, we take 100 000 samples of 10 000 users each. For each period of time we calculate the ARPU, variance of the revenue and the Conversion, obtaining the following histograms (Figure 2).

If we look at the scatterplots for each metric (Figure 3), we see that there seems to be a correlation between the two time periods for the ARPU and Conversion, whereas for the variance it is not as clear. Our hypothesis is that this could be due to some users with a considerable difference of revenue between the two time periods.

To confirm this we repeat the sampling removing the users with a difference larger than 1000USD between the two months. In this case, only 15 users have been taken out but we can clearly see the difference between the plots (Figure 4). Now, the correlation seems evident. Obviously, when analyzing the data all users must be considered and we are not able to remove the outliers, but this shows the effect a few users have when they are very different from the majority.

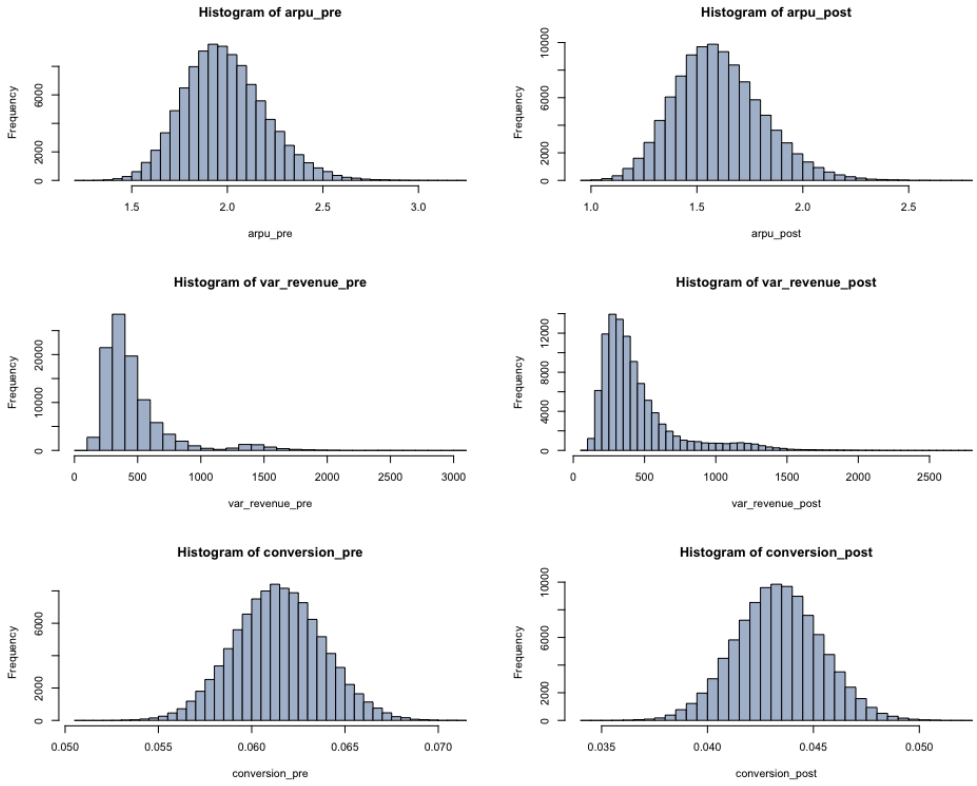

Figure 2: Histograms of the ARPU, variance of the revenue and Conversion

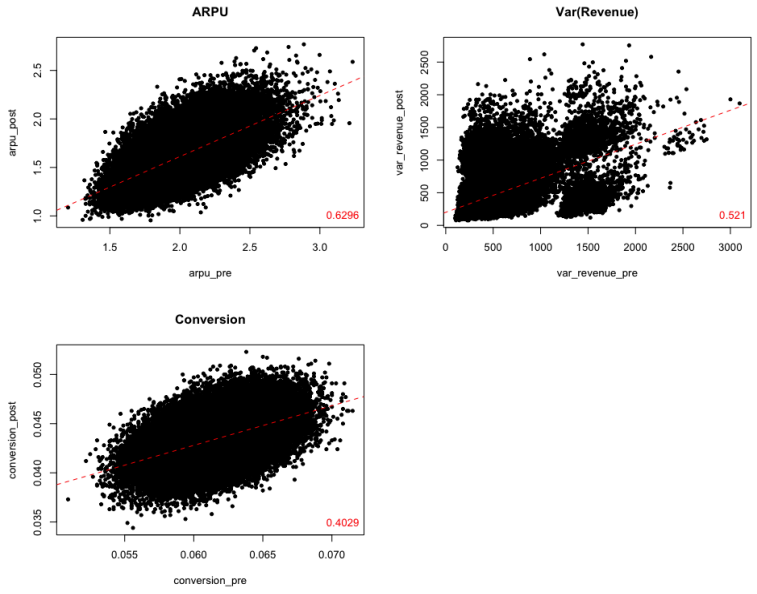

Figure 3: Scatterplots for the ARPU, variance of the revenue and Conversion in the pre and post periods

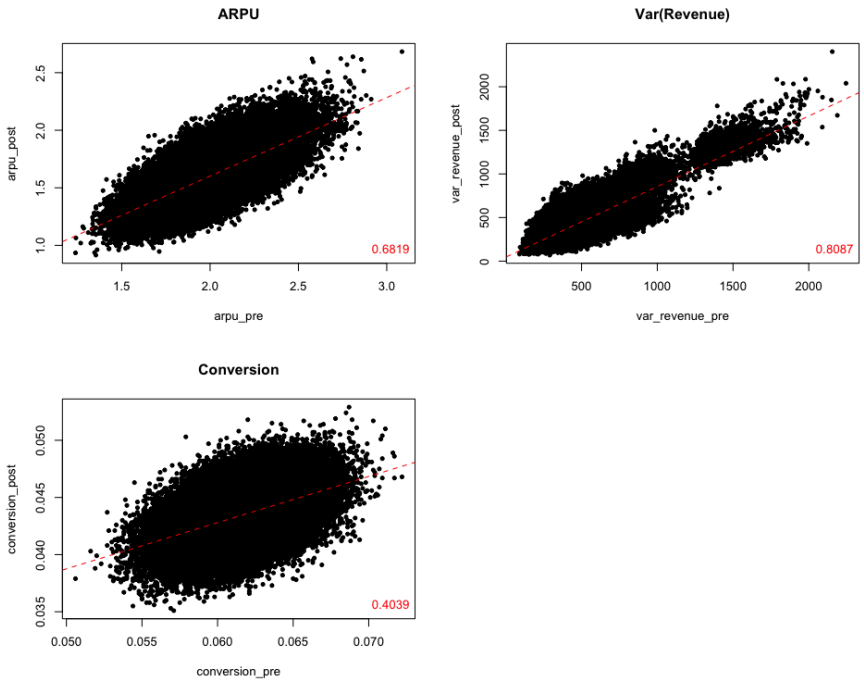

Figure 4: Scatterplots of the ARPU, variance of the revenue and Conversion after removing 15 users

Now, we take two different samples (A and B) and calculate the ARPU for each group as well as the delta for each time period ($\Delta_{pre} = ARPU^{B}_{pre} - ARPU^{A}_{pre}$). We repeat the process 100 000 times, taking two samples of 10 000 users each. In this case we see a clear correlation between the deltas, this means that if there is an overall difference between the two groups in the first period, there will likely be a comparable difference in the second period.

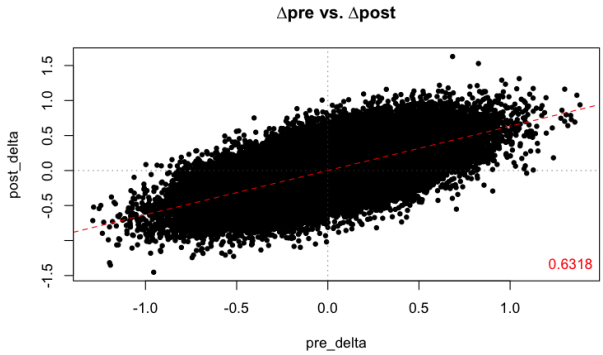

Figure 5: Scatterplot of Δ_{pre} vs. Δ_{post}

In this case, the results obtained for the other two games are very similar to what we just observed (plots can be found in the Annex A).

2. Existing analysis: understanding and reproduction

The next step of the process is to understand the existing analysis used for AB tests and be able to reproduce it; to then modify it, improve it and compare the results.

Apart from the report metrics, we will also calculate the value of the metric we are studying for each combination of {control, test} x {pre, post}, as well as the difference between groups for both periods of time, which we will refer to as pre delta and post delta.

2.1 Comparison to past AB tests

Once we have a reproduction of the code and before simulating AB tests, we are going to take data from real tests and see if we obtain the same results, just to check that the code is working correctly. We must take into account that when calculating the report metrics the results may slightly vary with each iteration, since there is a random sampling process involved. Therefore, we will repeat this step five times, to get an idea of how different the values we are obtaining are.

Firstly, we take a test (Test 2-1) that was active for two months, capturing around 870 000 users. In this case, our "post" data will contain the information for the whole duration of the test.

For the ARPU, the results obtained with the original processing are the following.

Group	Value	Uplift	Prob Beat Control	Uplift If Better	Uplift If Worse
0	$0.48	0.00%	0.00%	0.00%	0.00%
1	$0.43	-10.74%	0.05%	0.00%	-10.56%

Table 2: Report metrics obtained for the ARPU (Test 2-1)

And these are the results we obtained from our code.

Iter	Value (0)	Value (1)	Uplift	Prob Beat Control	Uplift If Better	Uplift If Worse
1	$0.48	$0.43	-10.70%	0.10%	0.00%	-10.66%
2	$0.48	$0.43	-10.70%	0.05%	0.00%	-10.64%
3	$0.48	$0.43	-10.70%	0.15%	0.00%	-10.53%
4	$0.48	$0.43	-10.70%	0.10%	0.00%	-10.53%
5	$0.48	$0.43	-10.70%	0.00%	0.00%	-10.45%

Table 3: Metrics obtained for the ARPU with our code (Test 2-1)

The values for the ARPU in both groups are practically the same (0.48$ and 0.43$), while the uplift varies a little (from -10.74% to -10.70%). For the values that change with the sampling, we originally had a P(beat control) = 0.05, and we obtained probabilities between 0 and 0.15. The uplift if better we obtained is quite close to 0 in all cases, and the uplift if worse has some differences in the first decimal position. Despite there being some slight differences, we can say that the results we obtained are quite similar to the original ones.

And these are the results for the Conversion.

Group	Value	Uplift	Prob Beat Control	Uplift If Better	Uplift If Worse
0	2.50%	0.00%	0.00%	0.00%	0.00%
1	2.46%	-1.77%	8.55%	0.06%	-1.77%

Table 4: Report metrics obtained for the Converison (Test 2-1)

Iter	Value (0)	Value (1)	Uplift	Prob Beat Control	Uplift If Better	Uplift If Worse
1	2.52%	2.48%	-1.73%	10.25%	0.07%	-1.79%
2	2.52%	2.48%	-1.73%	9.60%	0.06%	-1.80%
3	2.52%	2.48%	-1.73%	9.80%	0.05%	-1.77%
4	2.52%	2.48%	-1.73%	9.85%	0.06%	-1.79%
5	2.52%	2.48%	-1.73%	8.75%	0.06%	-1.79%

Table 5: Metrics obtained for the Conversion with our code (Test 2-1)

As we can see, we have very similar values for both groups, as well as for the uplift. But for the P(beat control) we see some slight differences, while the AB test analysis obtains a probability of 8.55%, the results from our analysis suggest slightly higher probabilities. For both the uplift if better and if worse, we obtained similar results to the ones in the report.

Now we do the same but with the data of another AB test (Test 2-2). In this case, the test was active for a longer period of time, around four months, and captured around 2 800 000 users. The results obtained are the following.

ARPU

Group	Value	Uplift	Prob Beat Control	Uplift If Better	Uplift If Worse
0	$1.24	0.00%	0.00%	0.00%	0.00%
1	$1.29	3.37%	89.05%	3.58%	-0.15%

Table 6: Report metrics obtained for the ARPU (Test 2-2)

Iter	Value (0)	Value (1)	Uplift	Prob Beat Control	Uplift If Better	Uplift If Worse
1	$1.31	$1.36	3.51%	89.55%	3.74%	-0.14%
2	$1.31	$1.36	3.51%	90.15%	3.74%	-0.11%
3	$1.31	$1.36	3.51%	90.65%	3.67%	-0.11%
4	$1.31	$1.36	3.51%	89.95%	3.71%	-0.13%
5	$1.31	$1.36	3.51%	88.75%	3.65%	-0.15%

Table 7: Metrics obtained for the ARPU with our code (Test 2-2)

Conversion

Group	Value	Uplift	Prob Beat Control	Uplift If Better	Uplift If Worse
0	3.42%	0.00%	0.00%	0.00%	0.00%
1	3.45%	0.92%	92.80%	0.95%	-0.02%

Table 8: Report metrics obtained for the Conversion (Test 2-2)

Iter	Value (0)	Value (1)	Uplift	Prob Beat Control	Uplift If Better	Uplift If Worse
1	3.46%	3.49%	1.00%	94.05%	0.99%	-0.02%
2	3.46%	3.49%	1.00%	94.35%	1.03%	-0.01%
3	3.46%	3.49%	1.00%	94.05%	1.02%	-0.01%
4	3.46%	3.49%	1.00%	94.60%	1.04%	-0.02%
5	3.46%	3.49%	1.00%	94.55%	1.02%	-0.01%

Table 9: Metrics obtained for the Conversion with our code (Test 2-2)

Now we can see that the differences between the report results and ours have accentuated. This could be due to some filters applied to the data in the original analysis; considering that this test spanned for a longer period of time and captured a lot more users, these minor differences could have been amplified. Still, we consider that our reproduction of the code is good and we can continue with the analysis.

2.2 AB tests simulation

Our next step is to sample from the game data to simulate an AB test. First of all, we simulate 10 000 AB tests that capture one million users each and plot the pre-test bias as a percentage. We do this for all three games and obtain the following histograms.

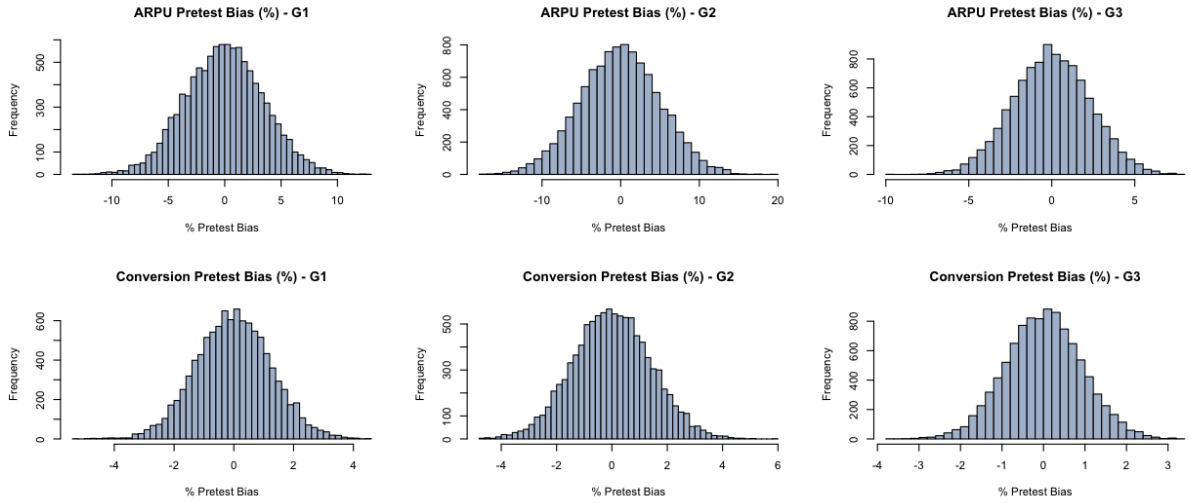

Figure 6: Histograms of the pretest bias (%) for ARPU and Conversion for each game

From these results, we can see the differences in the pretest bias between all games and also between the metrics we are studying. For example, we clearly see that the Conversion has a lower pretest bias than the ARPU; and G1 and G2 have higher values than G3 for both metrics. As a summary, we can say that 95% of the data are in the following intervals:

	G1	**G2**	**G3**
ARPU	[-6.8, 6.8] %	[-9.7, 9.9] %	[-4.6, 4.6] %
Conversion	[-2.4, 2.4] %	[-2.7, 2.7] %	[-1.7, 1.8] %

Table 10: Intervals of pretest bias (%) that contain 95% of the data for each game

We now focus on data from Game 1, also considering the information from one month as "pre" and the following month as "post" (test). We have observations for around 460 000 users, so we take samples of 100 000 users for both periods of time and randomly allocate each user in a group (control or test), then we apply our processing to obtain the metrics we want to study. We repeat the experiment 1 000 times for both the ARPU and Conversion. When comparing the probability to beat the control group to the pre and post deltas, we can see some interesting results. Clearly, the post delta has a very strong correlation with the probability to beat the control group, since we have only used the test data to calculate the metrics. The relation with the pre delta is not as clear, given that the pretest information is not used to get any results but, as we have seen above, both deltas are related.

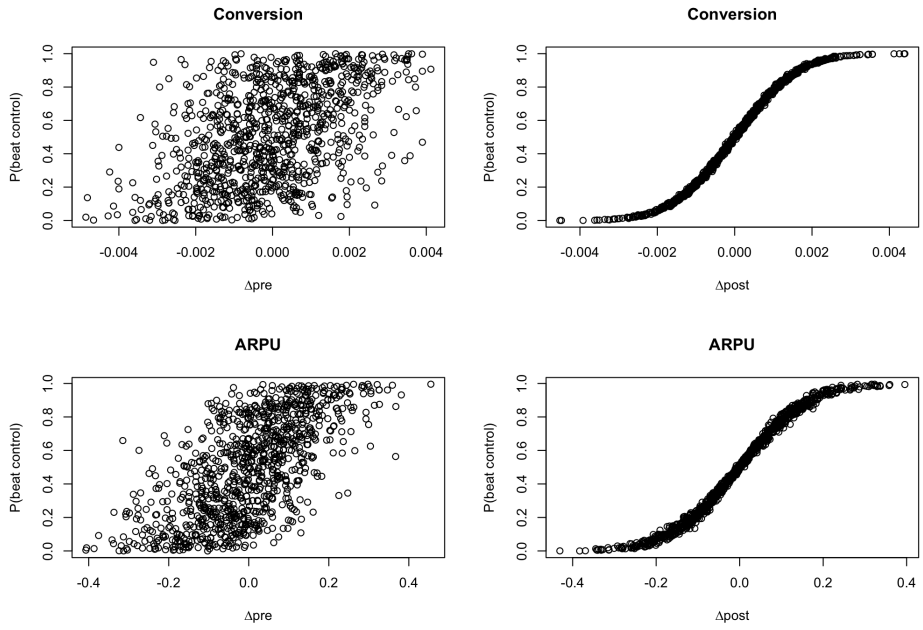

Figure 7: Scatterplots of the P(beat control) vs. Δ_{pre} and Δ_{post} for ARPU and Converison

Another interesting result are the histograms of the uplift and probability to beat the control group. In this case, we see that the P(beat control) is quite uniform, meaning that given one sample of the data, we can get any number between 0 and 1 as the P(beat control) with a very similar chance. For the uplift, on the other hand, we can see that the distribution is centered around 0, which makes sense given our data.

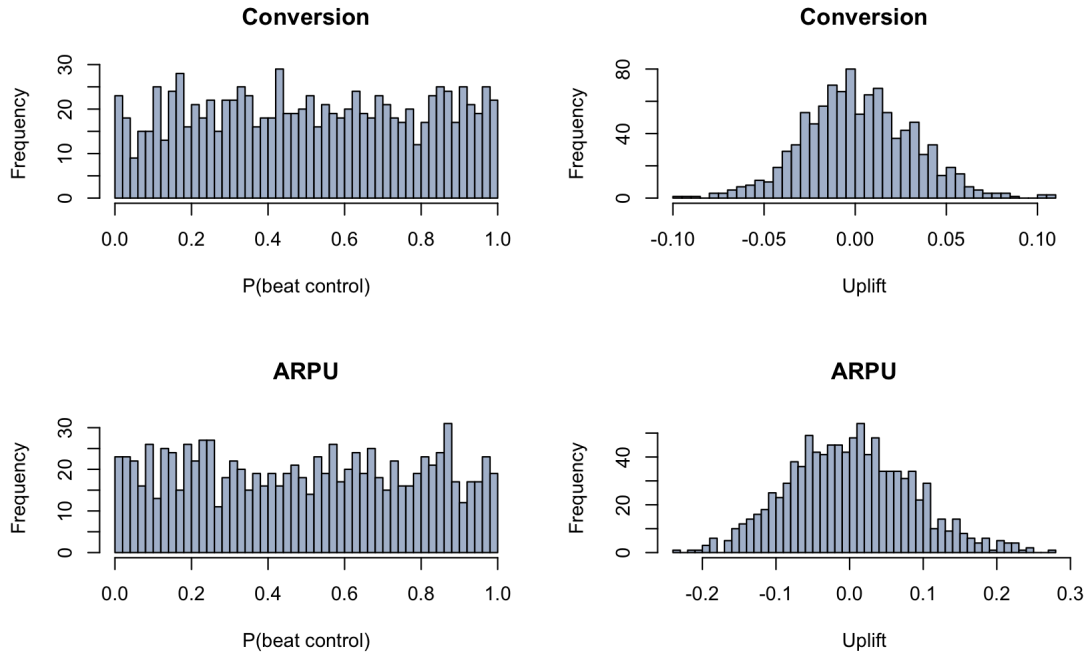

Figure 8: Histograms of P(beat control) and uplift for ARPU and Conversion

We now focus on the probability of beating the control group. Given that both groups come from the same population and are alike, we would think that the probability of one beating the other would be around 0.5, but instead we have a uniform distribution. To try to understand why this happens we will work with the Conversion. Since it is a proportion, the likelihood follows a Binomial distribution; and for the prior we use a non-informative Beta distribution. With this, we obtain a beta posterior distribution. Working with a specific random sample of the data, we have the following distributions for group A (control) and group B (test).

$$A \sim Binom(k = 24958, n = 50000)$$
$$B \sim Binom(k = 25051, n = 50000)$$

And when we apply the non-informative prior (Beta(α=1, β=1)) we obtain the following distributions.

$$A \sim Beta(\alpha = 24959, \beta = 25043)$$
$$B \sim Beta(\alpha = 25052, \beta = 24950)$$

Then, we randomly generate 2 000 values for the Conversion following these distributions (emulating the sampling performed in our processing) and create a new boolean variable (1 if A > B, 0 otherwise). When averaging this new variable we obtain a value of 0.279, meaning that in just 27.9% of the cases the value in group A is bigger than the one in group B.

We repeat the process but generating the values from distributions that are more balanced:

$$A \sim Beta(\alpha = 25000, \beta = 25000)$$
$$B \sim Beta(\alpha = 25000, \beta = 25000)$$

In this case, we obtain that in 50.4% of the cases the value of A is bigger than B. With this, we see that with just a small variation in our distributions there is a big change in the results and that the Beta distribution is quite sensitive to its parameters.

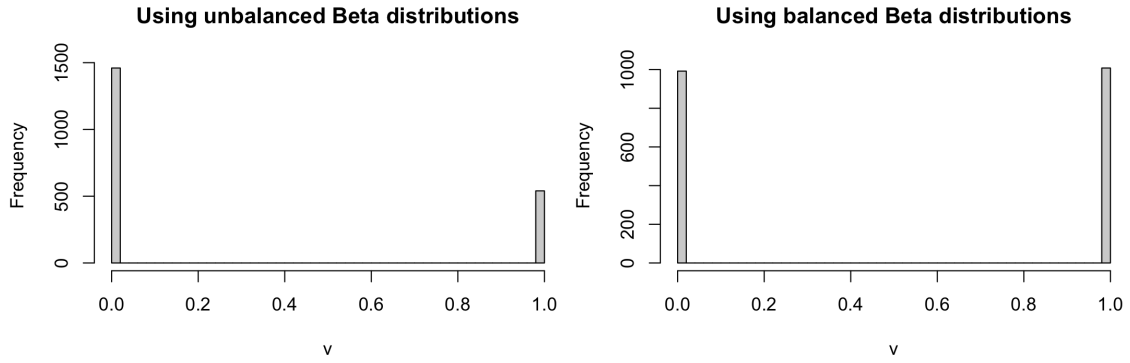

Figure 9: Comparison of results using balanced and unbalanced Beta distributions

In summary, we have seen that when doing an AB test, if the two groups come from the same population the probability of one group beating the other is uniformly distributed; which can be quite misleading, since we can obtain a probability of 0 or 1 when the two groups are very similar.

2.3 New metrics

As observed, the probability to beat the control group is quite misleading and unreliable, so we have three different options to face this problem. The first one would be to do nothing, which is what is happening now, and accept that the results may be wrong.

The second option is to consider a threshold for the difference between groups. That is, instead of considering P(A < B), work with P(A - B < Δ). This leaves a margin for error that can be chosen depending on the game or the metric we are working with. Since the processing for each metric is a little different we will start working with the Conversion.

We take both samples from the Binomial model with the Beta uninformative prior and compare the P(win control) we obtained with the P(win control with a X% margin), with X being 1, 2, 5 and 10. After repeating the experiment 1 000 times simulating AB Tests that capture 100 000 users, the results obtained are the following. The plots below show the relation between both probabilities, while the histograms show the distribution of these new metrics for each margin of error.

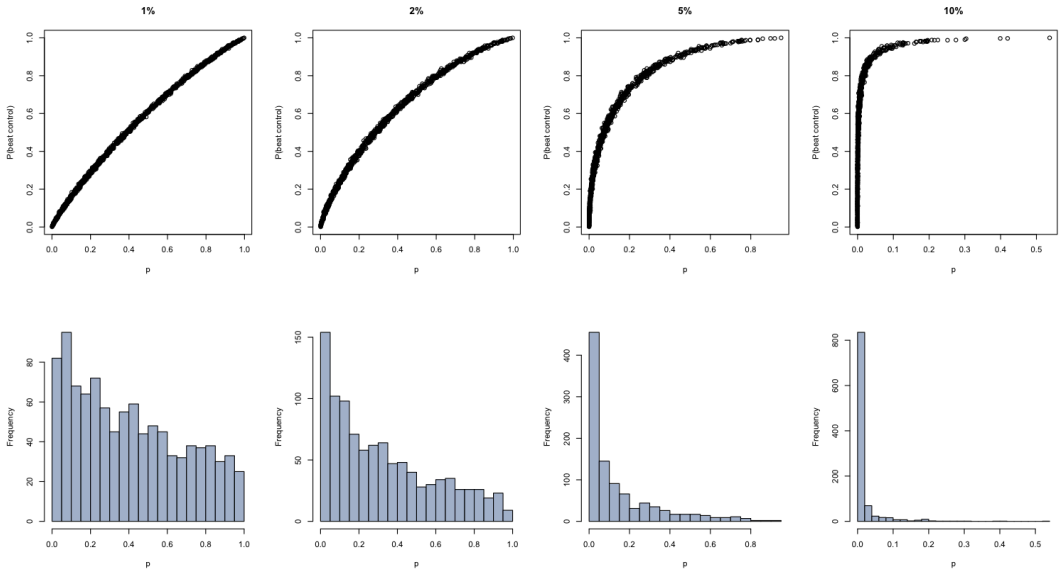

Figure 10: Results considering a margin for P(beat control) for Conversion

When repeating the experiment with the ARPU, the margin deltas chosen are 5%, 10%, 15% and 20%.

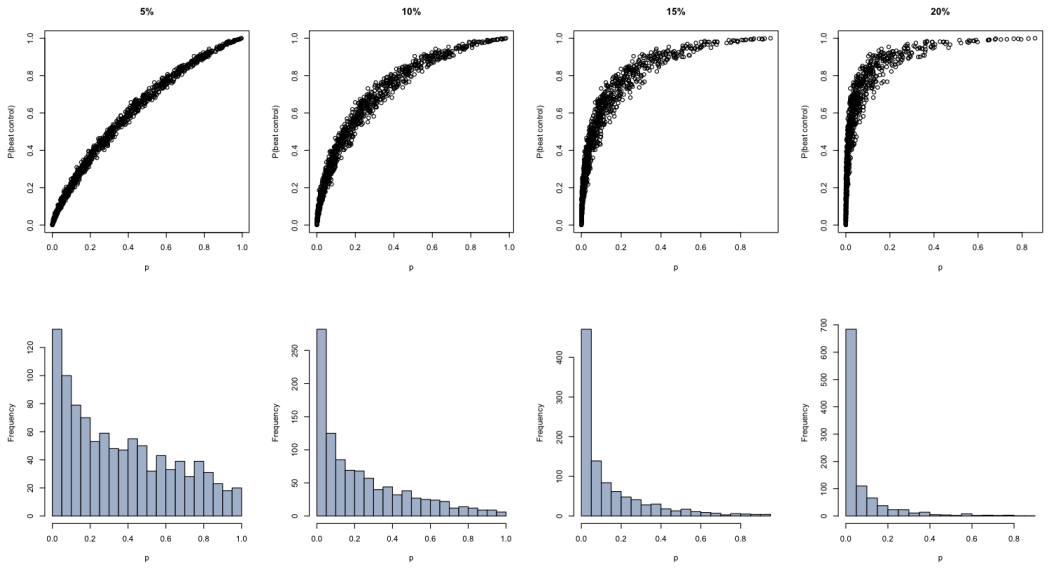

Figure 11: Results considering a margin for P(beat control) for ARPU

In both cases we can see some improvements, depending on the threshold chosen. But this delta can be very sensible to the specific test data we are working with and the metrics of interest. So this might not be the best alternative.

The third option is to consider the standard error of the estimator $\widehat{p}$ as a threshold. When working with the Conversion, since we use the Binomial distribution, this standard error is

$$SE = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

With n being the sampling size in our bayesian processing (2 000) and $\widehat{p}$ being the sampled probability for group 0 (control). We also calculate the difference between the samples from groups 0 and 1 and get a new variable based on the difference and error that is

- 0, if dif < -error
- 0.5, if -error ≤ dif ≤ error
- 1, if dif > error

Now, to calculate our new probability to beat the control group we calculate the mean of the new variable. Repeating the experiment 1 000 times, we can see that our new metric is centered around 0.5 and the results are not as misleading as with the original probability.

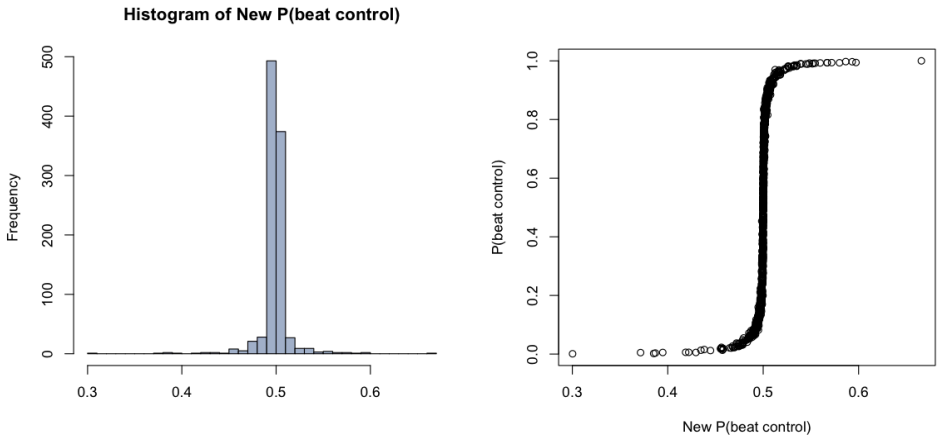

Figure 12: Results considering the standard error as a margin for the new P(beat control)

Up until now we have simulated AB tests that capture 100 000 users, but tests are often bigger and obtain data from a larger number of users. So the next step is to repeat the analysis for the Conversion with a bigger sample size. In this case, we work with 1 200 000 observations of data, randomly assigning each user to one of the two groups for each experiment.

When manually choosing a margin of error we see some improvements. However, it is quite tricky to choose the appropriate margin for each metric, test or game; so this is clearly not a good alternative.

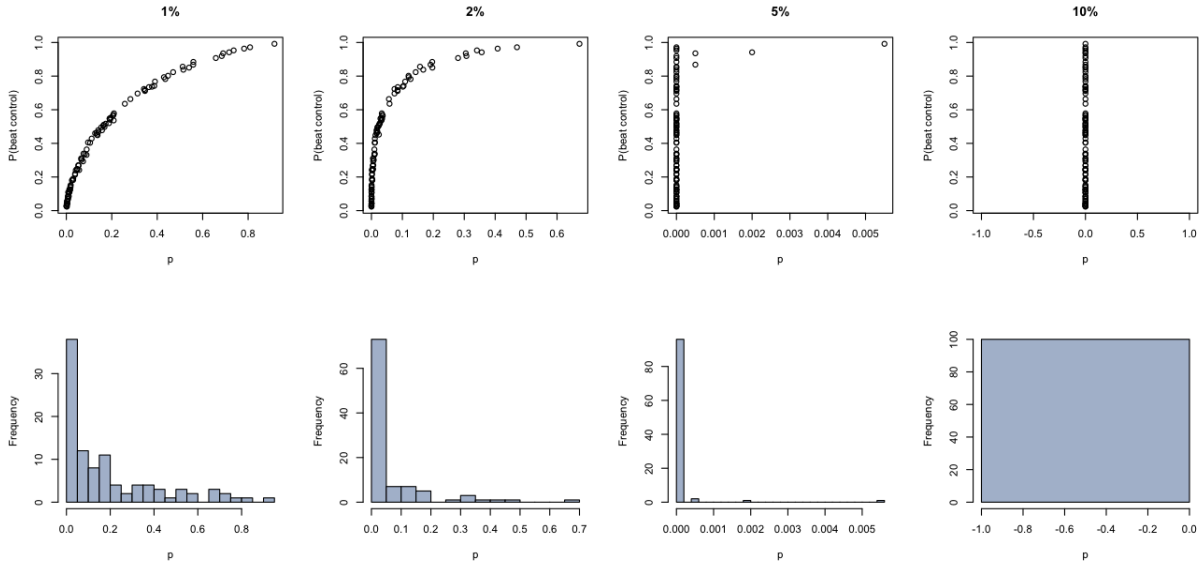

Figure 13: Results when considering a margin for P(beat control) for a larger sample

On the other hand, when considering the standard error as the margin and taking a larger sample, we see that all users fall into the interval [-SE, SE].

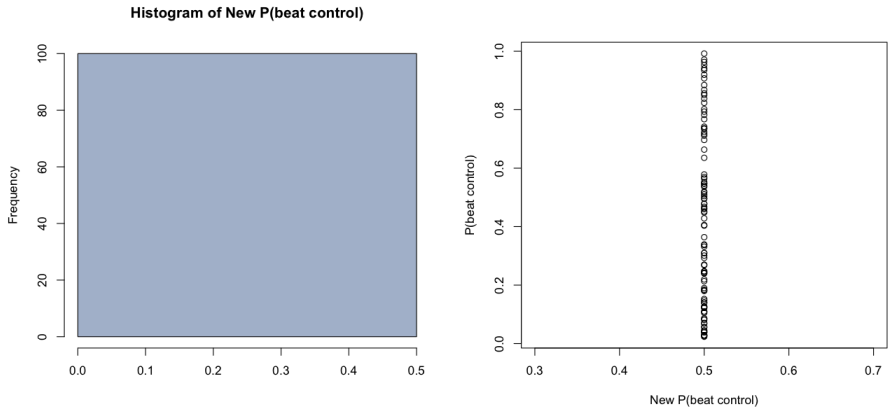

Figure 14: Results when considering a the error as margin for P(beat control) for a larger sample

In any case, we will continue exploring other ways to handle this metric.

3. Credible Intervals

When applying the current bayesian processing we obtain a credible interval (CI) for the uplift metric. If the interval is positive or negative we can say that one group is better than the other, but if zero is included in it the results are inconclusive. Sometimes, we only see the uplift of the probability to beat the control group, but we also have to look at the CI to extract some meaningful conclusions.

To see how often we might be wrong in our conclusions we do the following experiment: for different sample sizes we take a group of users, process their data and look at the CI obtained. We repeat the experiment 1 000 times, with sample sizes of 100k, 500k, 1M and 2M (this is the total sample size, so we divide the users into two groups of half the size). And we also modify the test group (B) to see if the results change depending on the difference between both groups. Apart from this, we also analyze the probability to beat the control group; considering its mean, Q1 and Q3.

If we look at the results for the ARPU, considering a confidence level of 95%, we can see how the percentage of inconclusive responses decreases when increasing the sample size, and also

when the difference between both groups gets bigger. This means that the bigger the sample size and the difference between groups, the easier it is to notice this contrast and say that group B is better.

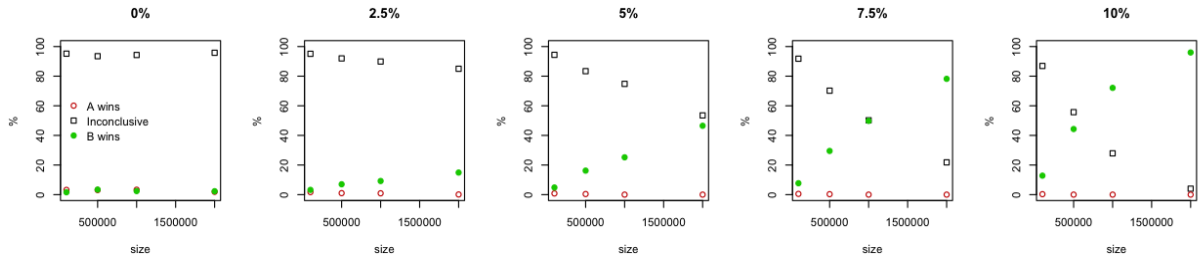

Figure 15: Evolution of percentages for a 95% confidence level (ARPU)

If we focus on the probability of beating the control group, we have previously seen that when both groups are equal it follows a uniform distribution. However, as the difference between groups and the sample size increases, the probability also grows.

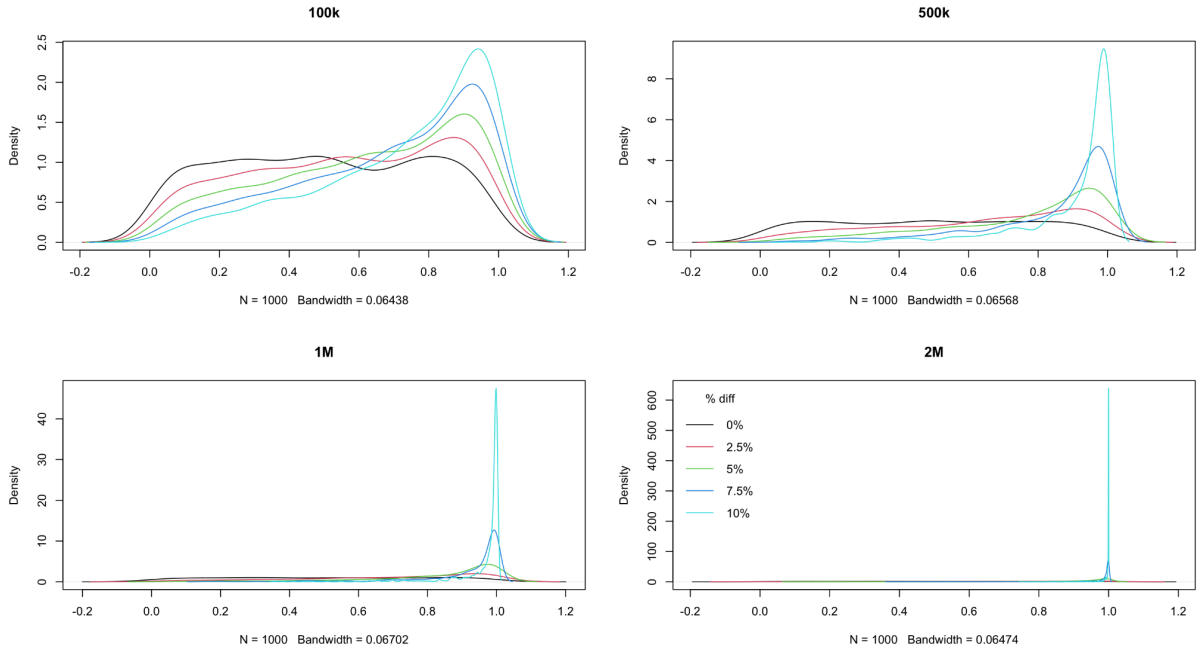

Figure 16: Evolution of P(beat control) for a 95% confidence level (ARPU)

We repeat this experiment for various confidence levels (95%, 90% and 85% – results can be found in the Annex B) and we can say that the error (column A wins) is bigger for smaller confidence. Moreover, when reducing the confidence the inconclusive percentage gets slightly smaller, and although the B wins percentage becomes somewhat bigger, the difference does not seem significant. All in all, the differences do not appear to be very meaningful when changing the confidence level.

Comparing the results for the Conversion with the previous ones, we can see a very similar trend, although the percentages for B wins grow faster. This means that it is significantly easier to detect when the test group wins. Additionally, it is easier to perceive these differences with smaller sample sizes.

As we can see in the following plot, the line for the 5% difference for a sample size of 500 000 users is higher than the line for a 10% difference for 100 000 users. Therefore, it is easier to discern a 5% difference with 500 000 users than a 10% difference with 100 000 users.

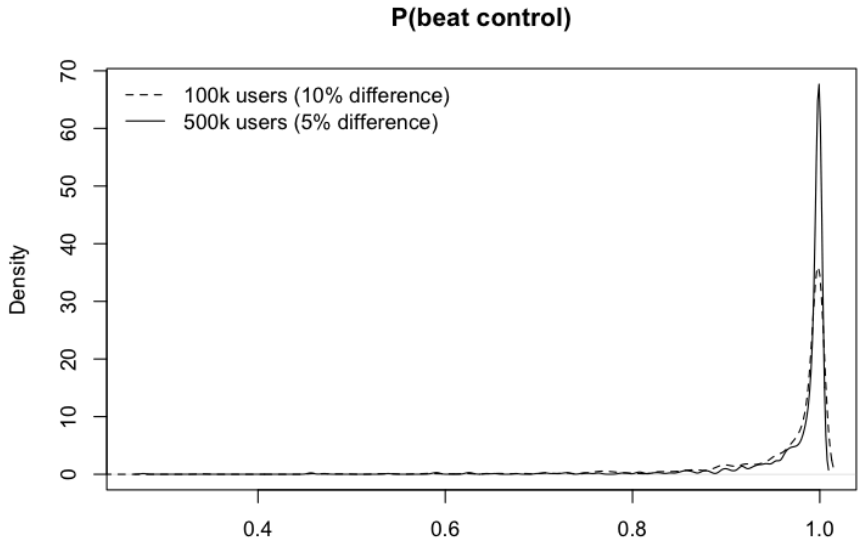

Figure 17: Comparison of P(beat control) for two specific cases (Conversion)

If we look at the following plot and compare it to the one we have done with the ARPU data, we see a very similar trend: the percentage of B winning grows when increasing the sample size or the difference between both groups. As we can see, this growth happens faster for the Conversion.

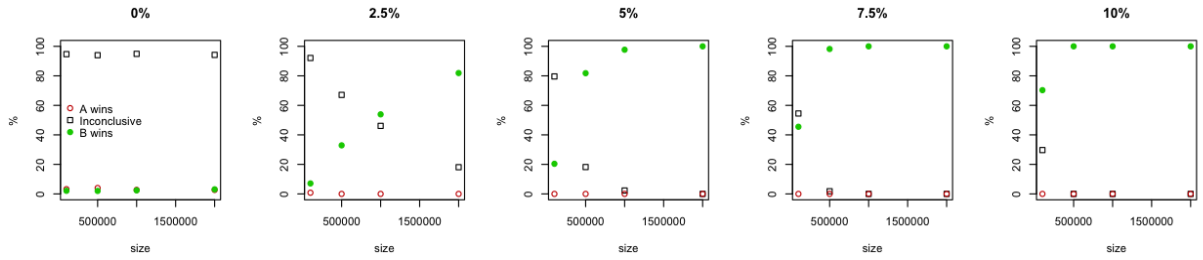

Figure 18: Evolution of percentages for a 95% confidence level (Conversion)

4. Other methods

Now we will try some new methods. First of all, we will try a mixture model that we think might be a better fit for the ARPU data. Then, we will also test some ways to reduce the bias considering the pretest information, since up until now we have only worked with data from the test (what we call "post" data) and completely disregarded the pretest.

4.1 Mixture

For now, we run the model for both groups of the post data and then obtain the report metrics. If we compare them with what we obtained with the current approach, we see that the general distributions are very similar. Although we do see that the uplift distribution for the current approach is more rounded and the P(beat control) for the new model is slightly more flat.

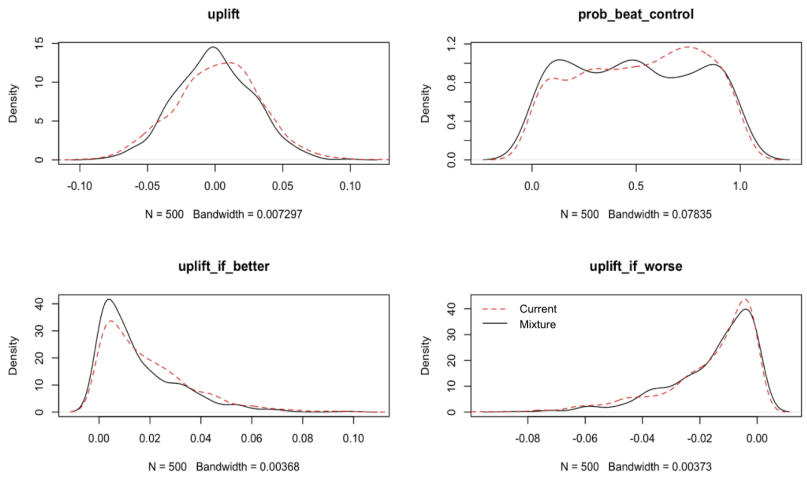

Figure 19: Comparison of report metrics with the current approach and mixture model

If we compare the correlation between the uplift and P(beat control) for both models we also see similar results, though for the new model we have more uplift values closer to zero. So it looks like this new model might be slightly better than the first one.

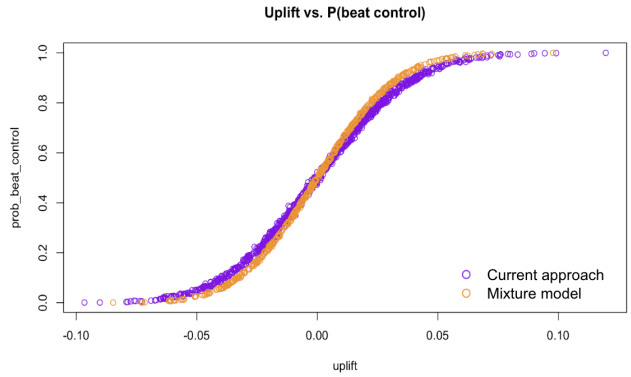

Figure 20: Comparison of the correlation between P(beat control) and uplift for the current approach and mixture model

We also compare the posterior distributions of both models with the original data. To clearly see the differences we transform the data applying a logarithm. If we focus on all the users, since the original processing works with a NormalGamma-Normal distribution and we have negative values, we apply the log-transformation to the X axis. However, if we focus only on paying users we can apply the transformation to all the data. It should be noted that the current approach works with positive and negative values for the revenue and transforms the negative values into non-paying users. On the other hand, the mixture model assumes that there are non-paying users. As we can see in the plots, it looks like the new approach is considerably better than the first one, although in the previous plots these differences are more difficult to appreciate.

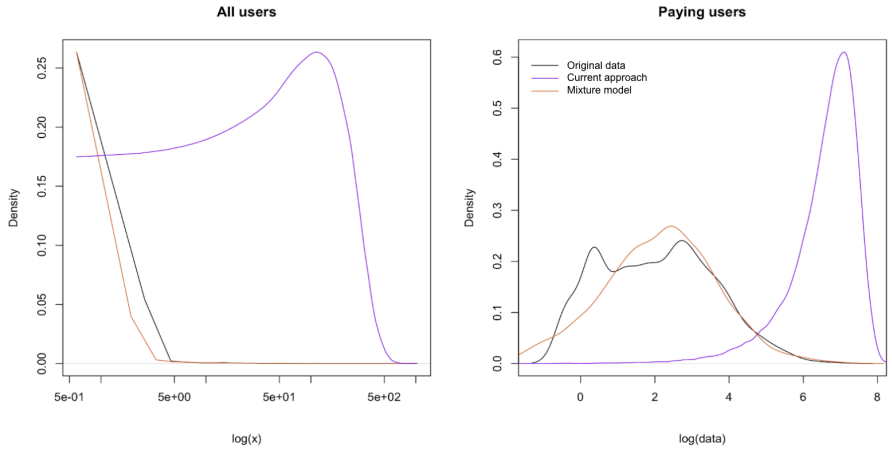

Figure 21: Comparison of posterior predictive distributions for the current approach and mixture model with the original data, applying a log-transformation

To check if the models can pick up on differences between groups, we manually increase the revenue of the data for the test group by 1% and 5% and run both models. When comparing the report metrics for both methods with and without the differences it looks like they both recognize them, although it seems like the mixture model might be slightly better.

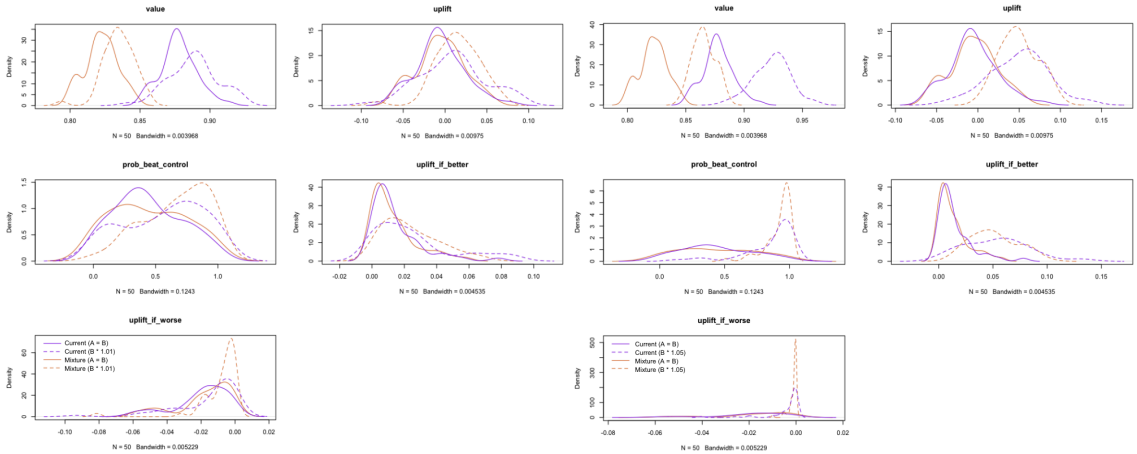

Figure 22: Comparison of report metrics for the current approach and mixture model, with equal groups and multiplying group B by 1.01 (left) and 1.05 (right)

We also see that the metric value is different for each model, so we study this by repeating the experiment fewer times (5, 10, 20 and 50) to see if there are any outliers involved in this change (Figure 23). As we can see, we obtain different results in all cases. This could be due to the differences in the way the values are calculated. As mentioned above, the current method samples from the corresponding distribution and obtains the values averaging these samples. On the other hand, the mixture model generates samples for the hyperparameters, that are later used to calculate the samples for the ARPU:

$$ARPU = exp(\mu + \frac{1}{2\tau})\cdot(1 - p)$$

Then, with the average of these samples we get the value for the report. Considering that the ARPU for the data before processing it is around 0.88, we see that these results might be misleading.

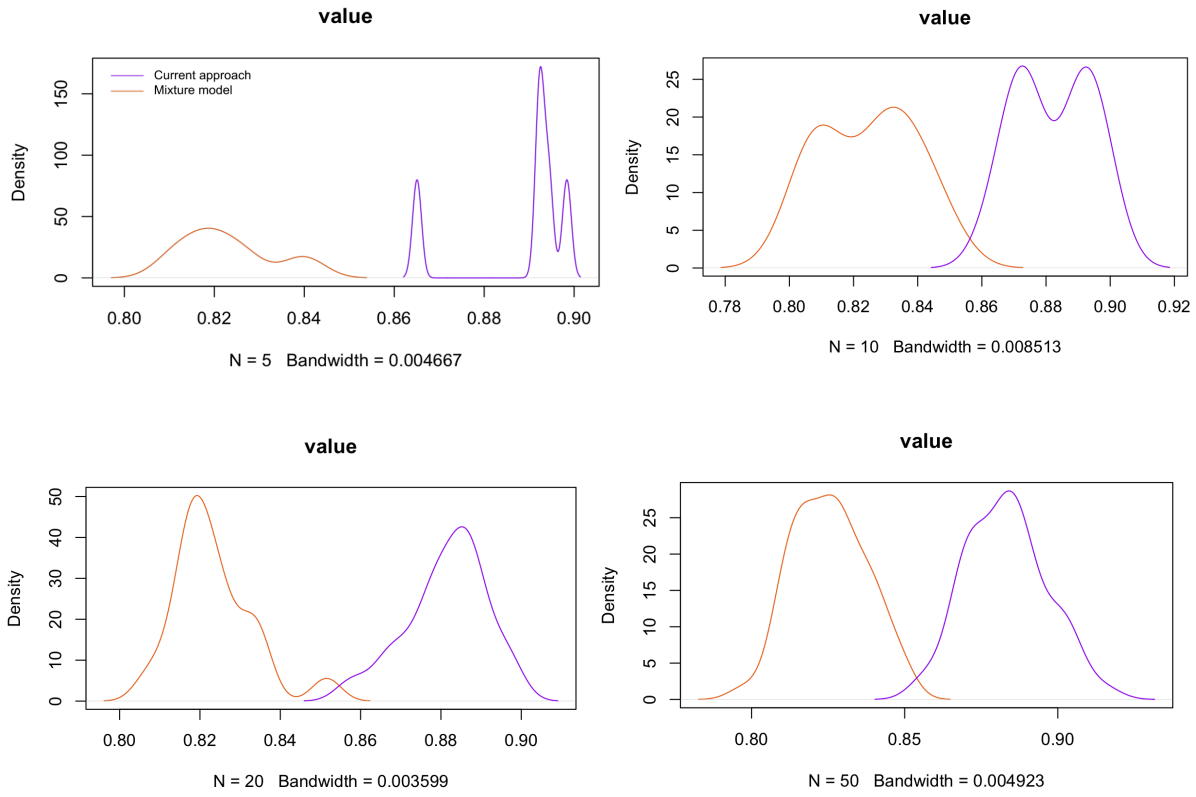

Figure 23: Comparison of values for ARPU obtained with the current approach and mixture model, for sample sizes of 5, 10, 20 and 50

Overall, we have seen that this new model might give us slightly better results for the uplifts. However, the value of the ARPU obtained for each group is somewhat different compared to the current approach and the original data; and although this difference might be a few cents, it is still important in our analysis. Another interesting outcome is that, despite the posterior distributions being very different, the overall results in the report are quite similar. Considering all this, and that the computation time is considerably higher for the mixture model, we have decided to keep the current processing for now, and explore other alternatives that might improve the results.

4.2 Bias reduction

Now we will explore three different methods to reduce the bias observed. These are known bias reduction techniques that we think might help us solve the problem.

4.2.1 CUPED

Firstly, we explore the CUPED methodology (Controlled-experiment Using Pre-Experiment Data) [1]. In this case, we don't analyze the post data, but a transformation of it that also takes into account the revenue from before the test:

$$\widehat{Y}_i = Y_i^{post} - (Y_i^{pre} - \overline{Y}^{pre}) \cdot \theta, \quad \text{with } \theta = \frac{cov(Y^{pre}, Y^{post})}{var(Y^{pre})}$$

We do 1 000 experiments with the ARPU using this method. If we look at the correlation plots between the report metrics and the deltas we can see some differences between the original model and this new method.

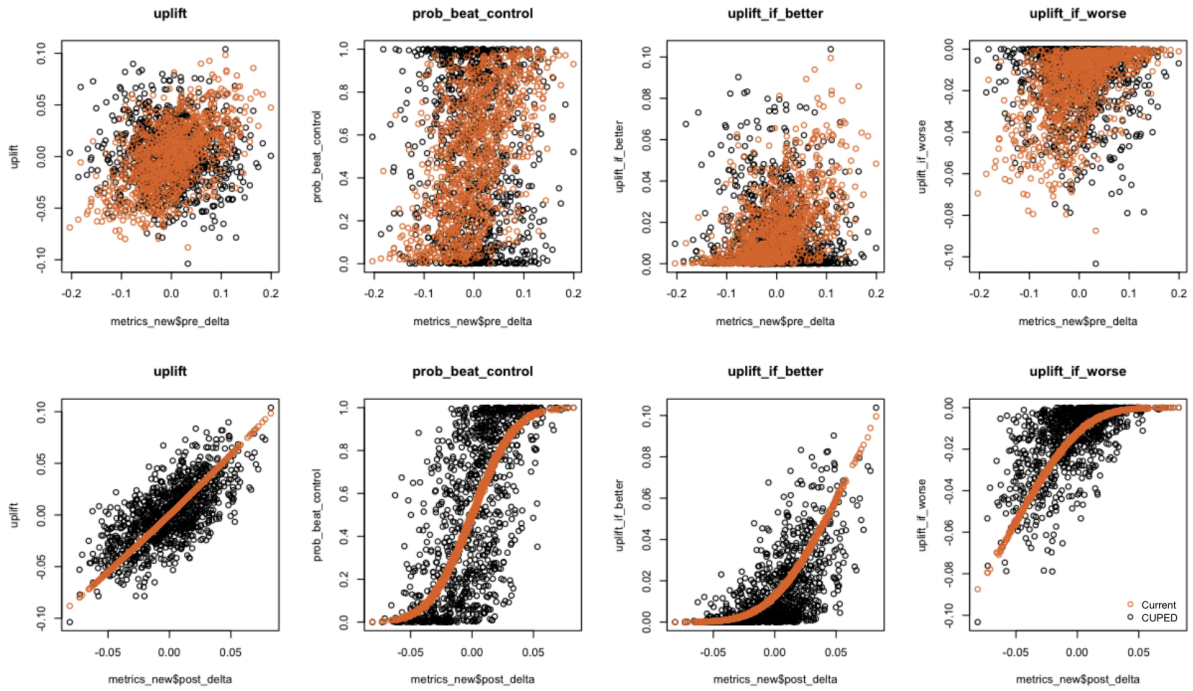

Figure 24: Comparison of report metrics for the current approach and CUPED

In particular, if we focus on the first plot and look at both sets of points separately, it looks like the correlation between the pretest delta and the report uplift disappears.

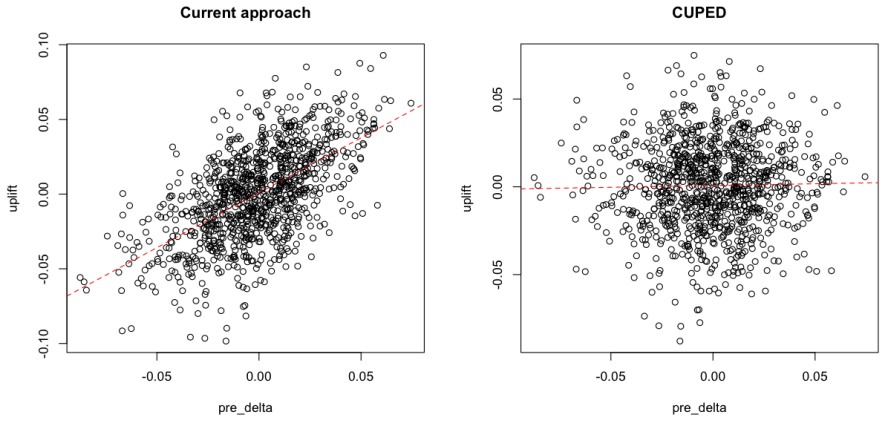

Figure 25: Correlation plots for the uplift vs. pretest bias
for the current approach and CUPED

4.2.2 Increment

Another hypothesis that we want to investigate is analyzing the difference between the test (post) and pretest data for each user. For this, we will look at the difference and apply a correction with the average pretest data so that it is equivalent to CUPED with θ = 1:

$$\widehat{Y}_i = Y_i^{post} - Y_i^{pre} + \overline{Y}^{pre}$$

Again, we compare the correlation between the deltas and the report metrics for this new method and the current approach. The results are quite similar to the ones we observed with CUPED, although in this case there is still a correlation between the uplift and pretest information.

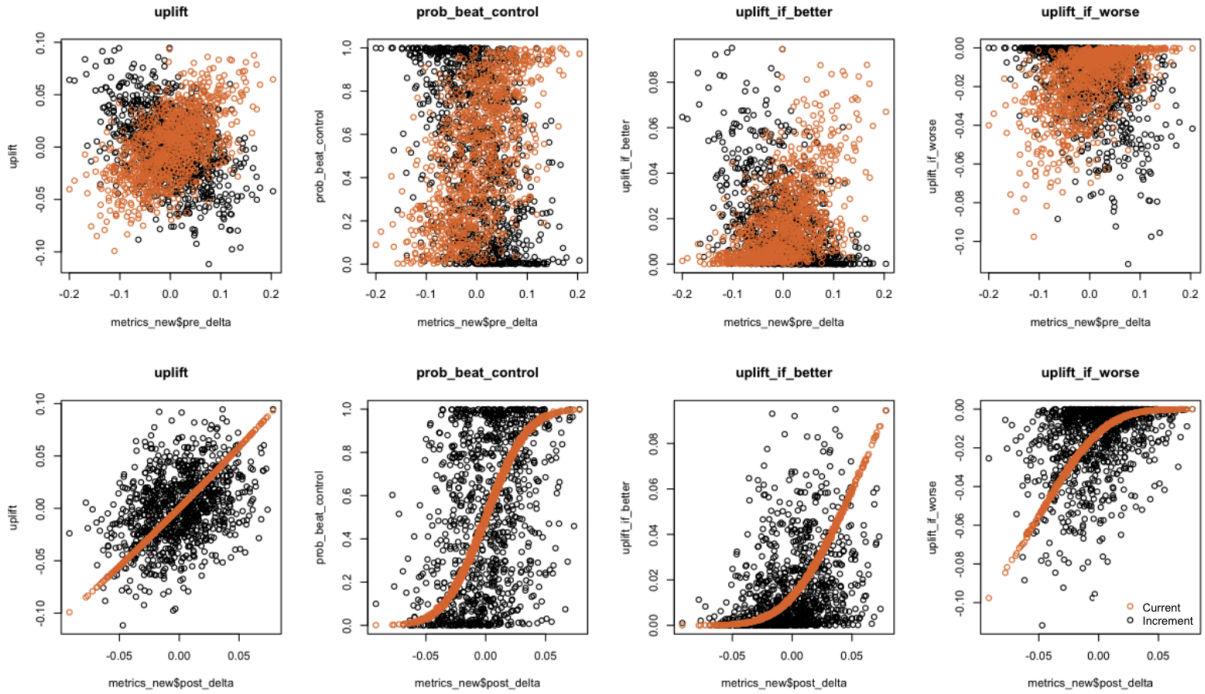

Figure 26: Comparison of report metrics for the current approach and Increment method

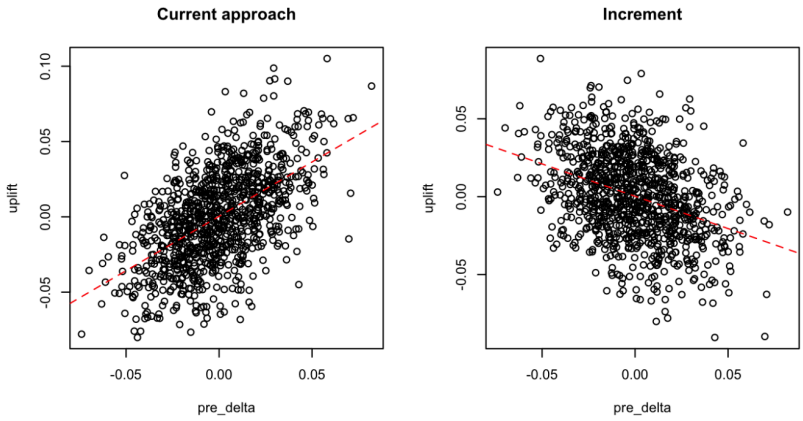

Figure 27: Correlation plots for the uplift vs. pretest bias for the current approach and Increment method

4.2.3 Test - pretest

Finally, we will subtract the pretest bias to the test metric for each user, only for the users in the test group. Namely,

$$\widehat{Y_i} = Y_i^{post} - \Delta_{pre}$$

In this case, we see that the correlation plot between the pretest bias and the uplift is very similar to the previous one for the Increment method.

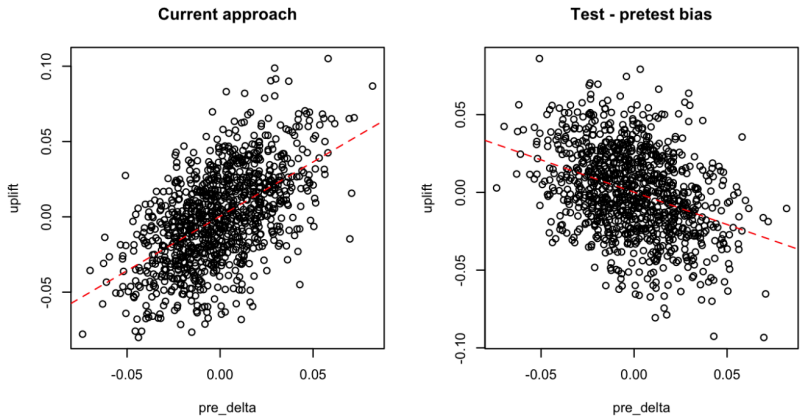

Figure 28: Correlation plots for the uplift vs. pretest bias for the current approach and Test - pretest method

5. Past AB tests

We now focus on past AB tests. When analyzing the data, we have some users who joined the game during the test, so we do not have pretest information for them. In this case, we transform the NAs in our pretest data into zeros. We process the data with all the methods we have seen up until now and compare the results with the ones obtained from the report.

- Current: current approach, only taking into account post data
- CUPED
- Increment: CUPED with θ = 1 (equal to working with the difference post - pre)
- Test - pretest: take post data and subtract the average pre-test bias (pre delta)

5.1 Test 1

Pre data: 30 days before entering test
Post data (test): 32 days
Total users: 1 374 004

ARPU:

Method	Value (0)	Value (1)	Uplift	CI (95%)		P(beat control)	Uplift if better	Uplift if worse
Report	1.37	1.39	1.78%	-2.91%	6.27%	78.35%	2.12%	-0.28%
Current	1.37	1.39	1.78%	-2.60%	6.29%	77.00%	2.03%	-0.29%
Test - pretest	1.37	1.42	4.13%	-0.67%	8.56%	96.10%	4.14%	-0.04%
Increment	1.37	1.43	4.12%	1.18%	7.29%	99.80%	4.18%	0.00%
CUPED	1.37	1.42	3.62%	0.88%	6.61%	99.20%	3.62%	0.00%

Table 11: Results obtained for ARPU using all methods (Test 5-1)

Conversion:

Method	Value (0)	Value (1)	Uplift	CI (95%)		P(beat control)	Uplift if better	Uplift if worse
Report	4.74%	4.74%	-0.07%	-1.80%	1.78%	46.95%	0.34%	-0.41%
Current	4.74%	4.74%	-0.07%	-1.89%	1.87%	48.30%	0.35%	-0.40%
Test - pretest	4.74%	4.79%	1.05%	-0.69%	2.89%	86.50%	1.09%	-0.06%
Increment	4.73%	4.78%	1.05%	-0.64%	3.01%	86.80%	1.10%	-0.06%
CUPED	4.73%	4.76%	0.67%	-1.07%	2.54%	76.75%	0.79%	-0.12%

Table 12: Results obtained for Conversion using all methods (Test 5-1)

As expected, the two first rows are very similar, given that the current method is a reproduction of the code used to generate the report. But if we compare this to the other methods we can observe some differences. For both the ARPU and the Conversion we obtain the lowest uplifts with this original bayesian processing, as well as the lowest P(beat control). While we get the highest values for the uplifts when we consider the test data minus the pretest bias and the difference between post and pre data (increment). This difference is more noticeable for the Conversion, since the uplift in the report is negative and with the other methods it's positive.

Another interesting metric we can look at is the credible interval for the uplift. For the ARPU results, we see that using both the report and test minus pretest methods, we obtain intervals that include 0, making the test non-significant. While for the other two methods the intervals are above 0, which means that the improvement of the test group compared to the control group is significant. So we can see how we can extract very different conclusions depending on the method we are using.

For the Conversion results, the differences are not as noticeable since all intervals include 0. In this case, looking at the CIs we would extract the same conclusion for all methods, but looking at the uplifts the interpretations are very different.

5.2 Test 2

Pre data: 30 days before entering test
Post data (test): 107 days
Total users: 2 631 129

ARPU:

Method	Value (0)	Value (1)	Uplift	CI (95%)		P(beat control)	Uplift if better	Uplift if worse
Report	1.29	1.25	-2.99%	-8.46%	2.27%	14.20%	0.21%	-3.11%
Current	1.29	1.25	-2.99%	-8.03%	2.21%	12.75%	0.17%	-3.21%
Test - pretest	1.29	1.28	-0.80%	-5.97%	4.79%	40.35%	0.82%	-1.48%
Increment	1.28	1.27	-0.81%	-4.67%	4.08%	35.00%	0.56%	-1.41%
CUPED	1.27	1.27	-0.23%	-4.72%	4.16%	44.90%	0.80%	-1.07%

Table 13: Results obtained for ARPU using all methods (Test 5-2)

Conversion:

Method	Value (0)	Value (1)	Uplift	CI (95%)		P(beat control)	Uplift if better	Uplift if worse
Report	3.55%	3.57%	0.45%	-0.92%	2.05%	72.65%	0.60%	-0.13%
Current	3.55%	3.56%	0.50%	-1.12%	1.89%	71.95%	0.59%	-0.13%
Test - pretest	3.55%	3.57%	0.70%	-0.86%	2.22%	81.20%	0.80%	-0.08%
Increment	3.54%	3.57%	0.70%	-0.92%	2.08%	80.40%	0.75%	-0.08%
CUPED	3.55%	3.57%	0.59%	-0.87%	2.25%	77.45%	0.71%	-0.10%

Table 14: Results obtained for Conversion using all methods (Test 5-2)

In this case, for the ARPU results, we also see that the report obtains the lowest uplift and the lowest probability to beat the control group. And we get the highest values for these metrics using CUPED. However, all the uplift values are negative and the CIs include 0, so the conclusions we would extract are all very similar for all methods.

If we now take a look at the Conversion results we see how, again, the lowest uplift and P(beat control) are obtained with the current method. Although in this case the highest values come from the test minus pretest and increment methodologies. Just like with the ARPU, the conclusions we can extract using each method would be quite similar in this case since all uplifts are positive and all CIs include 0, making the test non-significant.

6. Credible Intervals and Pretest Bias

Our last experiment is very similar to the one we have previously done with credible intervals. We also take a million users, divide them into two equal groups and repeat the experiment 5 000 times. In this case, however, we do not modify any of the groups and instead separate each sample by the pretest bias observed:

$$\Delta_{pre} = \bar{y}_{pre}^{test} - \bar{y}_{pre}^{control}$$

We process each sample with all the methods we are studying and count the percentage of cases where the control group wins (CI < 0), the test group wins (CI > 0) or the results are inconclusive ($0 \in$ CI). In this case, since both groups are equal, we expect the majority of the iterations to fall into the inconclusive column and consider the other cases as the error. For each method we calculate the error made and analyze the results for different metrics.
We start with the ARPU and consider the following pretest bias buckets:

- $\Delta \leq -0.04$
- $-0.04 < \Delta \leq -0.015$
- $-0.015 < \Delta \leq 0.015$
- $0.015 < \Delta \leq 0.04$
- $\Delta > 0.04$

It should be taken into account that we have users with no pretest information. In this case, we transform the NAs in our data into zeros. For now, we study three different time periods, considering one month as pretest and the following month as the test period:

Period	Pre	Post
1	Dec 2024	Jan 2025
2	Feb 2025	Mar 2025
3	Aug 2024	Sep 2024

Table 15: Time periods considered

For each combination of game and time period we obtain a table as the following, having the error for each pretest bias group and then calculating the total error, taking into account the number of samples within each bucket. We should note that CUPED has very similar results for all pretest buckets, due to the lack of correlation with the Δ_{pre} found above, but when obtaining the total error this does not seem to make a significant difference; while the other approaches work well with small pretest bias but considerably worse when it is large.

Game 1 - Period 1					
	Current	CUPED	Test-pretest	Increment	n
$\Delta \leq -0.04$	14.6%	5.2%	10.2%	11.3%	479
$-0.04 < \Delta \leq -0.015$	3.6%	4.8%	3.7%	4.0%	1070
$-0.015 < \Delta \leq 0.015$	2.0%	5.3%	1.9%	1.9%	1842
$0.015 < \Delta \leq 0.04$	4.2%	5.5%	5.1%	4.9%	1092
$\Delta > 0.04$	12.0%	6.6%	12.6%	14.3%	517
Error	5.06%	5.36%	4.89%	5.19%	

Table 16: Error for each method and pretest bias bucket
for Game 1 and Period 1, and overall error

Then, we compare the errors for all the games and time periods considered. As we can see, the Test - pretest method has the smallest error for all cases except one, and has considerably low error values for Game 3.

		Current	CUPED	Test - pretest	Increment
G1	Period 1	5.06%	5.36%	4.89%	5.19%
	Period 2	5.42%	5.37%	2.26%	5.02%
	Period 3	5.04%	5.17%	3.39%	5.31%
G2	Period 1	5.57%	5.48%	6.84%	4.91%
	Period 2	5.44%	5.68%	2.00%	5.28%
	Period 3	5.77%	5.82%	3.73%	5.69%
G3	Period 1	5.79%	4.99%	0.70%	4.49%
	Period 2	5.62%	5.20%	0.38%	5.58%
	Period 3	5.21%	5.04%	0.45%	5.38%

Table 17: Errors for ARPU, for each method, time period and game

For the Conversion metric the experiment is very similar, studying the same time periods. In this case, however, the pretest bias buckets are the following:

- $\Delta \leq -0.0004$
- $-0.0004 < \Delta \leq -0.00015$
- $-0.00015 < \Delta \leq 0.00015$
- $0.00015 < \Delta \leq 0.0004$
- $\Delta > 0.0004$

Here, it is clear that CUPED has the smallest error in all cases; and for G1 and G2, the Test - pretest and Increment methods have the largest errors. Again, we see how we obtain very small values for G3.

		Current	CUPED	Test - pretest	Increment
G1	Period 1	5.96%	4.21%	8.21%	8.24%
	Period 2	5.51%	3.89%	7.43%	7.51%
	Period 3	5.49%	4.25%	8.81%	8.92%
G2	Period 1	4.95%	4.14%	8.01%	7.91%
	Period 2	5.19%	3.70%	7.20%	6.80%
	Period 3	5.24%	3.93%	7.50%	7.13%
G3	Period 1	4.61%	1.53%	3.12%	2.97%
	Period 2	5.22%	1.85%	2.74%	2.48%
	Period 3	4.74%	1.38%	2.33%	2.15%

Table 18: Errors for Conversion, for each method, time period and game

Our next step is to analyze the results for the Daily ARPU (or DARPU), which is the average revenue per user divided by the number of days considered. That is, for the test data, we divide the ARPU by the number of days the test has been active. For the pretest information, considering a given pretest period; if the user has been registered for longer, we divide the ARPU by the number of days considered as pretest, but if the user registered during this period, we divide their ARPU by the number of days since registration. That is:

$$DARPU_{pre} = ARPU_{pre} / \min(days\_pretest, days\_from\_register)$$

In addition, when separating by pretest bias buckets, we will consider the pretest bias as a percentage:

$$pretest\ bias\ \% = \frac{\Delta_{pre}}{\bar{y}_{pre}} \cdot 100 = \frac{\bar{y}_{pre}^{test} - \bar{y}_{pre}^{control}}{\bar{y}_{pre}} \cdot 100$$

So the buckets considered are the following:

- $\Delta \leq -5\%$
- $-5\% < \Delta \leq -2\%$
- $-2\% < \Delta \leq 2\%$
- $2\% < \Delta \leq 5\%$
- $\Delta > 5\%$

For now, we will repeat each experiment considering the Daily ARPU for our pretest data and both the ARPU and DARPU for the test data, to see if there are any differences. Moreover, we

will remove the NGUs for some experiments and transform the pretest bias for the Test - pretest as follows

$$\text{pretest'} = \text{pretest} * \text{days\_test} / \text{days\_pretest}$$

We will simulate a test that started on March 1st 2025, and consider 60 days for the pretest period and 30 days for the test. When looking at the general results (with NGUs and without modifying the pretest bias), it looks like working with the Test minus pretest method is the best option when considering the test data as ARPU, but the worst option if we are working with Daily ARPU for the test. When removing the NGUs from the analysis, Test - pretest also appears to be the best method.

	Game	Post data	Current	CUPED	Test - pretest	Increment
	G1	DARPU	5.34%	6.09%	10.77%	5.66%
		ARPU	5.38%	5.43%	4.98%	5.31%
	G2	DARPU	4.92%	4.80%	8.20%	5.27%
		ARPU	5.63%	5.34%	4.91%	5.74%
- No NGUs	G2	DARPU	5.37%	5.32%	8.89%	5.26%
		ARPU	5.19%	5.64%	4.41%	5.06%
	G3	DARPU	5.12%	5.21%	0.76%	5.54%
		ARPU	5.00%	5.34%	4.52%	4.94%
- No NGUs - Modify pretest	G2	DARPU	4.92%	5.31%	2.22%	5.29%
		ARPU	5.21%	4.82%	4.62%	5.19%
	G3	DARPU	5.29%	6.02%	0.72%	5.46%
		ARPU	5.86%	5.12%	5.70%	5.77%

Table 19: Errors for DARPU, for each method and game

Finally, we analyze the Conversion again, but now removing the NGUs. In this case, the buckets considered are:

- $\Delta \leq -3\%$
- $-3\% < \Delta \leq -1.5\%$
- $-1.5\% < \Delta \leq 1.5\%$
- $1.5\% < \Delta \leq 3\%$
- $\Delta > 3\%$

If we also modify the pretest bias in the Test - pretest method, these are the results obtained. Just like above, we simulate an AB test that starts on the 1st of March 2025, take a million users to then separate them into two groups and repeat the process 5 000 times. In this case, we consider 60 days of pretest and 30 days of test; except for G3, where we consider 75 days of pretest and 75 days of test to have enough data to perform the analysis. We will not do the experiment for G1 removing the NGUs since we would need a much bigger time period to have enough data. As we can see, we get a slightly smaller error when removing the NGUs, except for the case when applying the current processing to G3 data and the Increment to G2. In any case, we obtain the smallest errors with each method for G3.

Game	Days pre	Days test	NGUs	Current	CUPED	Test - pretest*	Increment
G1	60	30	Keep	5.20%	4.21%	4.77%	10.66%
G2			Keep	5.55%	4.54%	4.87%	10.74%
			Remove	4.85%	3.11%	4.14%	14.47%
G3	75	75	Keep	5.08%	1.68%	3.37%	3.48%
			Remove	5.35%	1.39%	2.75%	2.79%

* Pretest bias modified

Table 20: Errors for Conversion, for each method and game, with and without NGUs

If we redo this experiment but without modifying the pretest bias in the Test - pretest method, we get bigger errors for this method for G1 and G2, although the results for G3 are very similar.

Game	Days pre	Days test	NGUs	Current	CUPED	Test - pretest	Increment
G1	60	30	Keep	5.45%	4.19%	10.79%	10.65%
G2			Keep	5.35%	3.97%	10.68%	10.54%
			Remove	4.81%	3.01%	13.17%	13.48%
G3	75	75	Keep	5.12%	1.98%	3.49%	3.52%
			Remove	5.16%	1.39%	2.88%	2.88%

Table 21: Errors for Conversion, for each method, time period and game

Conclusions

After all this work, we can extract some meaningful results. We have seen that the data has very heavy-tailed distributions, which can be a problem when analyzing it. In particular, users with the highest revenues, since large differences between the groups before a test lead to a significant pretest bias that must be considered.

We have also seen that the current analysis method has some problems, such as the probability of beating the control group following a uniform distribution when both groups are equal. This can be quite misleading when analyzing an AB test. For this, looking at the CI for the uplift might be a better option to decide if the test is significant or not.

Moreover, the current distribution used for the ARPU is not a very good fit for the data, although the results seem to be quite reliable. After trying a new model for this (a mixture model), we decided to keep the original processing for now.

We have also explored some bias reduction techniques to reduce the bias while including the pretest information in our analysis. For the ARPU, the best method seems to be subtracting the pretest bias to the test data for each user in the test group (Test - pretest method). However, CUPED seems to be the best method for the Conversion metric. In addition, we have seen that using CUPED the correlation between the pretest bias and the uplift disappears, so we get very similar errors regardless of the pretest bias. Thus, for higher values of the pretest bias, this method works very well. However, for values closer to 0 it adds some variance and the error is larger compared to other methods. We have also analyzed the impact of NGUs and we have seen that removing these users from all the analysis we obtain a smaller error, although we lose a significant amount of information along the way, since oftentimes users captured for a test have not played the game before.

Finally, although we have seen some very interesting results with this project, there are some aspects that require further investigation. So the next steps would be to do a deeper analysis on each method to better understand the results, especially focusing on the CUPED methodology. Additionally, it would be very beneficial to keep working with the NGUs and study how to handle them.

Bibliography

[1] Deng, A., Xu, Y., Kohavi, R., & Walker, T. (2013, February). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the sixth ACM international conference on Web search and data mining* (pp. 123-132).

Annex A. Sampling plots

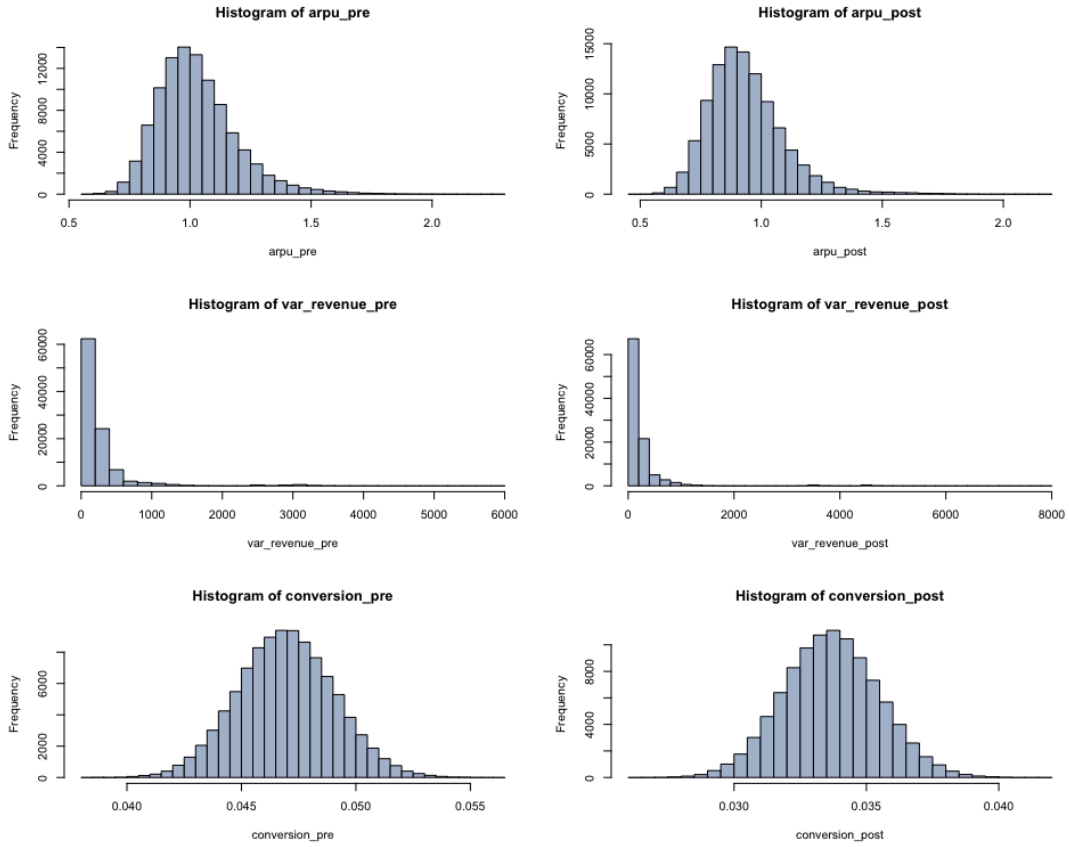

Figure A-1: Histograms of the ARPU, variance of the revenue and Conversion for Game 2

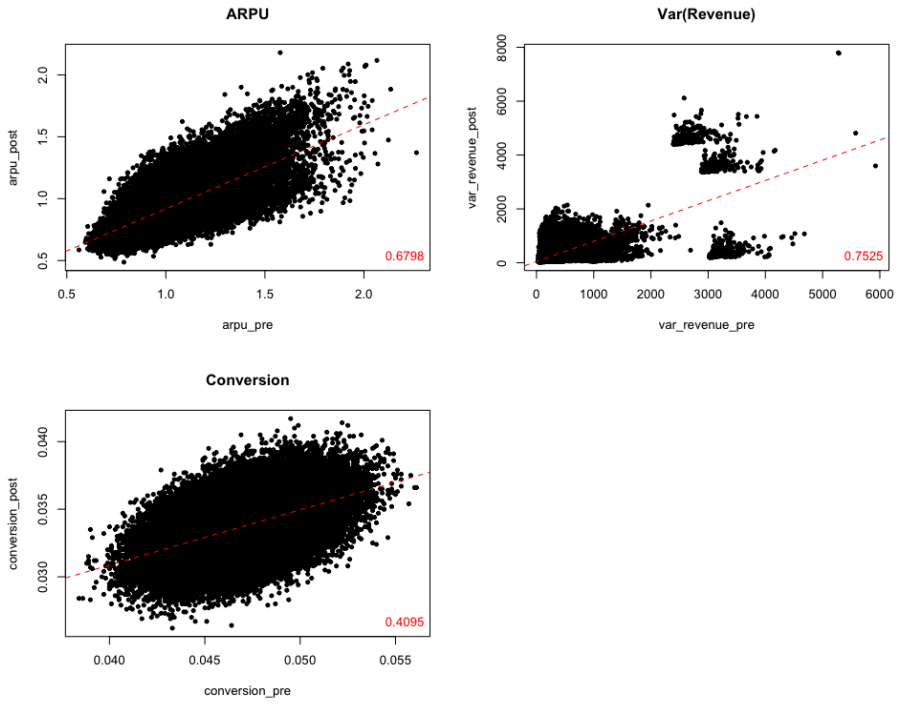

Figure A-2: Scatterplots for the ARPU, variance of the revenue and Conversion in the pre and post periods (Game 2)

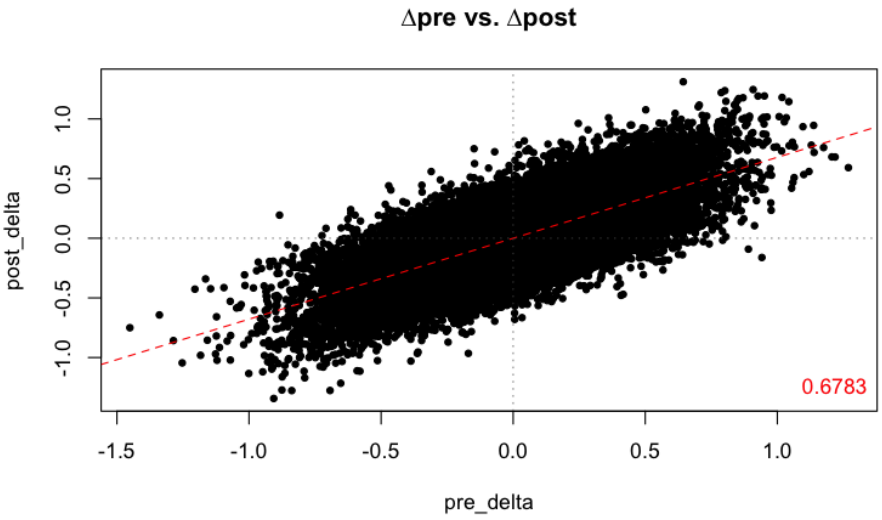

Figure A-3: Scatterplots of the ARPU, variance of the revenue and Conversion after removing the users with the bigger differences (Game 2)

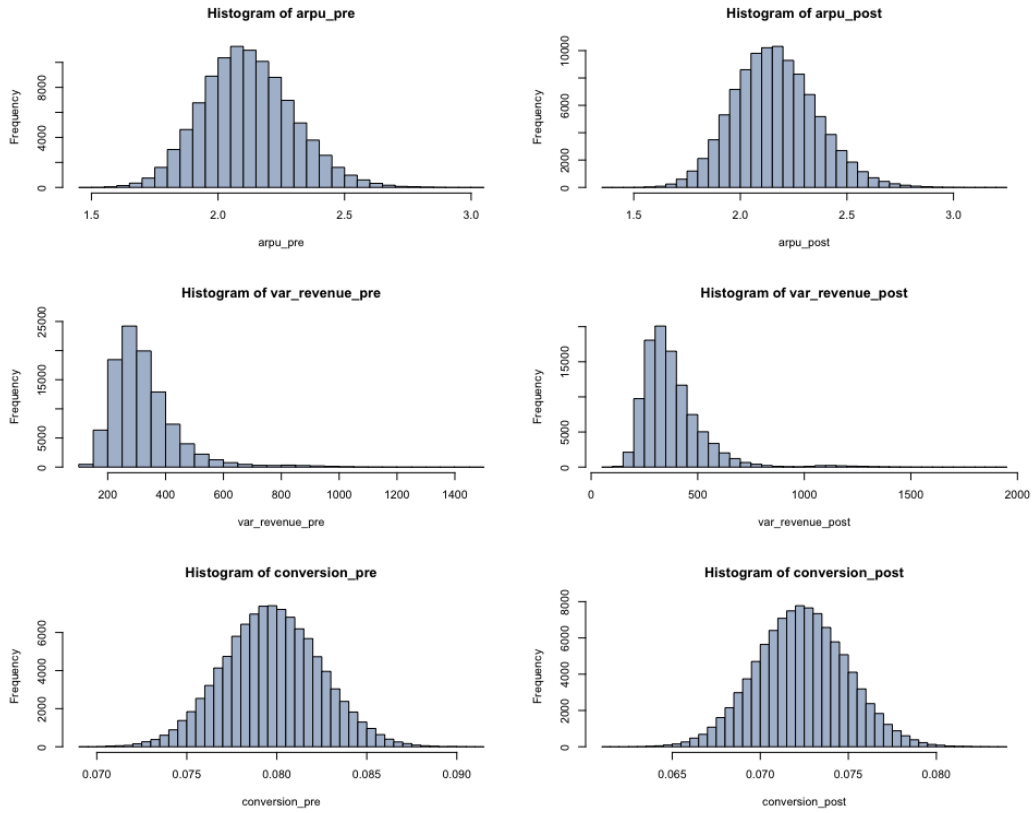

Figure A-4: Histograms of the ARPU, variance of the revenue and Conversion for Game 3

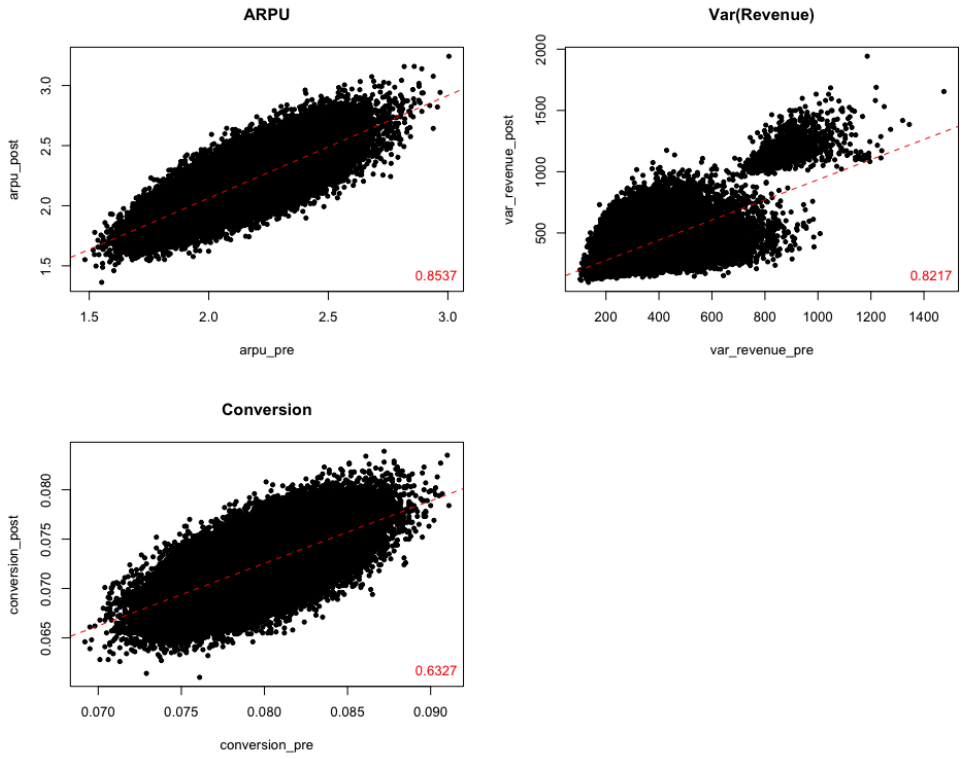

Figure A-5: Scatterplots for the ARPU, variance of the revenue and Conversion in the pre and post periods (Game 3)

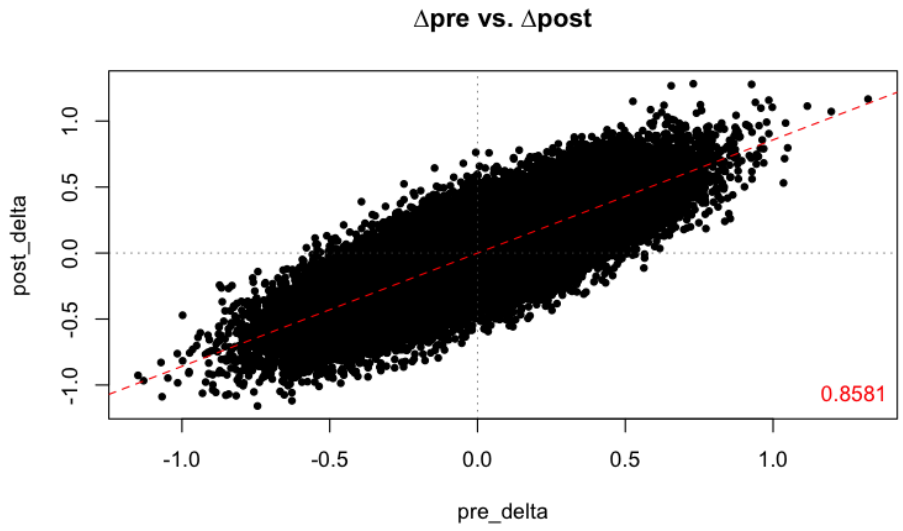

Figure A-6: Scatterplots of the ARPU, variance of the revenue and Conversion after removing the users with the bigger differences (Game 3)

Annex B. Credible Intervals tables

		CI (95%)			P(beat control)			
Total sample size	**%diff between groups**	**A wins**	**Inconclusive**	**B wins**	**Q1**	**Mean**	**Q3**	**Q3 - Q1**
100k	0.0%	3.5	94.4	2.1	0.26	0.51	0.77	0.51
	2.5%	2.1	94.7	3.2	0.33	0.57	0.83	0.50
	5.0%	1.0	92.8	6.2	0.41	0.63	0.88	0.47
	7.5%	0.4	90.7	8.9	0.49	0.68	0.92	0.43
	10%	0.1	87.4	12.5	0.57	0.73	0.94	0.37
500k	0.0%	2.5	95.3	2.2	0.26	0.50	0.75	0.49
	2.5%	1.1	92.5	6.4	0.44	0.64	0.88	0.44
	5.0%	0.3	84.9	14.8	0.62	0.75	0.95	0.33
	7.5%	0.0	72.0	28.0	0.78	0.84	0.98	0.20
	10%	0.0	54.9	45.1	0.88	0.91	0.99	0.11
1M	0.0%	2.8	94.4	2.8	0.25	0.50	0.76	0.51
	2.5%	0.1	90.5	9.4	0.51	0.69	0.92	0.41
	5.0%	0.1	73.3	26.6	0.75	0.83	0.98	0.23
	7.5%	0.0	49.5	50.5	0.91	0.92	0.99	0.08
	10%	0.0	26.4	73.6	0.98	0.97	0.99	0.01
2M	0.0%	3.1	94.5	2.4	0.26	0.50	0.75	0.49
	2.5%	0.1	84.1	15.8	0.61	0.75	0.95	0.34
	5.0%	0.0	52.8	47.2	0.89	0.91	0.99	0.10
	7.5%	0.0	19.5	80.5	0.98	0.98	1.00	0.02
	10%	0.0	3.9	96.1	0.99	0.99	1.00	0.01

Table B-1: Results for ARPU (95% confidence)

		CI (90%)			P(beat control)			
Total sample size	**% diff between groups**	**A wins**	**Inconclusive**	**B wins**	**Q1**	**Mean**	**Q3**	**Q3 - Q1**
100k	0.0%	6.0	90.4	3.6	0.22	0.49	0.74	0.52
	2.5%	3.7	90.6	5.7	0.30	0.55	0.81	0.51
	5.0%	2.5	88.0	9.5	0.38	0.61	0.87	0.49
	7.5%	1.1	84.6	14.3	0.46	0.67	0.90	0.44
	10%	0.7	80.9	18.4	0.54	0.72	0.93	0.39
500k	0.0%	4.9	90.6	4.5	0.26	0.50	0.75	0.49
	2.5%	1.9	86.9	11.2	0.45	0.64	0.87	0.42
	5.0%	0.4	76.6	23.0	0.62	0.75	0.95	0.33
	7.5%	0.1	59.9	40.0	0.78	0.84	0.98	0.20
	10%	0.0	43.6	56.4	0.89	0.91	0.99	0.10
1M	0.0%	4.5	90.2	5.3	0.25	0.5	0.75	0.50
	2.5%	0.8	83.5	15.7	0.5	0.68	0.91	0.41
	5.0%	0	62.3	37.7	0.75	0.83	0.98	0.23
	7.5%	0	39.2	60.8	0.9	0.92	1	0.10
	10%	0	16.7	83.3	0.97	0.97	1	0.03
2M	0.0%	6.7	88.2	5.1	0.25	0.5	0.76	0.51
	2.5%	0.2	74.5	25.3	0.61	0.75	0.95	0.34
	5.0%	0	41.7	58.3	0.89	0.91	1	0.11
	7.5%	0	13.1	86.9	0.98	0.98	1	0.02
	10%	0	2.1	97.9	1	1	1	0.00

Table B-2: Results for ARPU (90% confidence)

		CI (85%)			P(beat control)			
Total sample size	**%diff between groups**	**A wins**	**Inconclusive**	**B wins**	**Q1**	**Mean**	**Q3**	**Q3 - Q1**
100k	0.0%	8.3	85.3	6.4	0.27	0.51	0.76	0.49
	2.5%	5.3	84.6	10.1	0.35	0.57	0.82	0.47
	5.0%	3.5	80.8	15.7	0.43	0.63	0.87	0.44
	7.5%	2.2	78.2	19.6	0.51	0.68	0.91	0.40
	10%	0.9	73.2	25.9	0.60	0.74	0.94	0.34
500k	0.0%	8.4	85.3	6.3	0.23	0.50	0.76	0.53
	2.5%	3.5	79.5	17.0	0.40	0.63	0.89	0.49
	5.0%	1.8	67.8	30.4	0.59	0.75	0.95	0.36
	7.5%	0.5	48.9	50.6	0.76	0.84	0.98	0.22
	10%	0.1	34.9	65.0	0.87	0.90	0.99	0.12

Table B-3: Results for ARPU (85% confidence)

		CI (95%)			P(beat control)			
Total sample size	**%diff between groups**	**A wins**	**Inconclusive**	**B wins**	**Q1**	**Mean**	**Q3**	**Q3 - Q1**
100k	0.0%	3.6	94.1	2.3	0.27	0.50	0.74	0.47
	2.5%	0.9	90.9	8.2	0.48	0.67	0.90	0.42
	5.0%	0.2	77.5	22.3	0.71	0.80	0.97	0.26
	7.5%	0	58.2	41.8	0.87	0.89	0.99	0.12
	10%	0	35.3	64.7	0.96	0.95	1	0.04
500k	0.0%	3.2	94.7	2.1	0.21	0.47	0.72	0.51
	2.5%	0.2	76.2	23.6	0.71	0.81	0.97	0.26
	5.0%	0	28.5	71.5	0.97	0.97	1	0.03
	7.5%	0	2.6	97.4	1	1	1	0.00
	10%	0	0.2	99.8	1	1	1	0.00
1M	0.0%	2.8	94	3.2	0.25	0.5	0.76	0.51
	2.5%	0	52.1	47.9	0.89	0.91	1	0.11
	5.0%	0	3.6	96.4	1	1	1	0.00
	7.5%	0	0	100	1	1	1	0.00
	10%	0	0	100	1	1	1	0.00
2M	0.0%	2.4	95.7	1.9	0.24	0.49	0.73	0.49
	2.5%	0	24.2	75.8	0.98	0.97	1	0.02
	5.0%	0	0	100	1	1	1	0.00
	7.5%	0	0	100	1	1	1	0.00
	10%	0	0	100	1	1	1	0.00

Table B-4: Results for Conversion (95% confidence)

		CI (90%)			P(beat control)			
Total sample size	**%diff between groups**	**A wins**	**Inconclusive**	**B wins**	**Q1**	**Mean**	**Q3**	**Q3 - Q1**
100k	0.0%	5.4	89.7	4.9	0.27	0.5	0.74	0.47
	2.5%	1.1	84.3	14.6	0.49	0.67	0.90	0.41
	5.0%	0.2	69.1	30.7	0.72	0.81	0.97	0.25
	7.5%	0.0	44.3	55.7	0.88	0.90	0.99	0.11
	10%	0.0	23.0	77.0	0.96	0.95	1	0.04
500k	0.0%	6.0	89.1	4.9	0.25	0.50	0.74	0.49
	2.5%	0.2	62.8	37.0	0.75	0.83	0.98	0.23
	5.0%	0.0	15.6	84.4	0.98	0.97	1	0.02
	7.5%	0.0	1.5	98.5	1	1	1	0.00
	10%	0.0	0.0	100.0	1	1	1	0.00
1M	0.0%	5.1	89.0	5.9	0.26	0.50	0.73	0.47
	2.5%	0.0	39.0	61.0	0.89	0.91	0.99	0.10
	5.0%	0.0	1.5	98.5	1	1	1	0.00
	7.5%	0.0	0.0	100.0	1	1	1	0.00
	10%	0.0	0.0	100.0	1	1	1	0.00
2M	0.0%	5.2	90.5	4.3	0.26	0.50	0.74	0.48
	2.5%	0.0	14.6	85.4	0.98	0.97	1	0.02
	5.0%	0.0	0.0	100.0	1	1	1	0.00
	7.5%	0.0	0.0	100.0	1	1	1	0.00
	10%	0.0	0.0	100.0	1	1	1	0.00

Table B-5: Results for Conversion (90% confidence)