Saint-Petersburg State University

Nenakhov Ivan Vladimirovich

Final qualifying work
Methods of A/B testing of web forms

Education level: bachelor’s degree

Education program 02.03.02 «Fundamental computer science and information technology»

Scientific supervisor:
Candidate of Engineering, Associate Professor
Blekanov I.S.

Reviewer:
Candidate of Physics and Mathematics, Associate Professor
Korovkin M.V.

Saint-Petersburg

2024

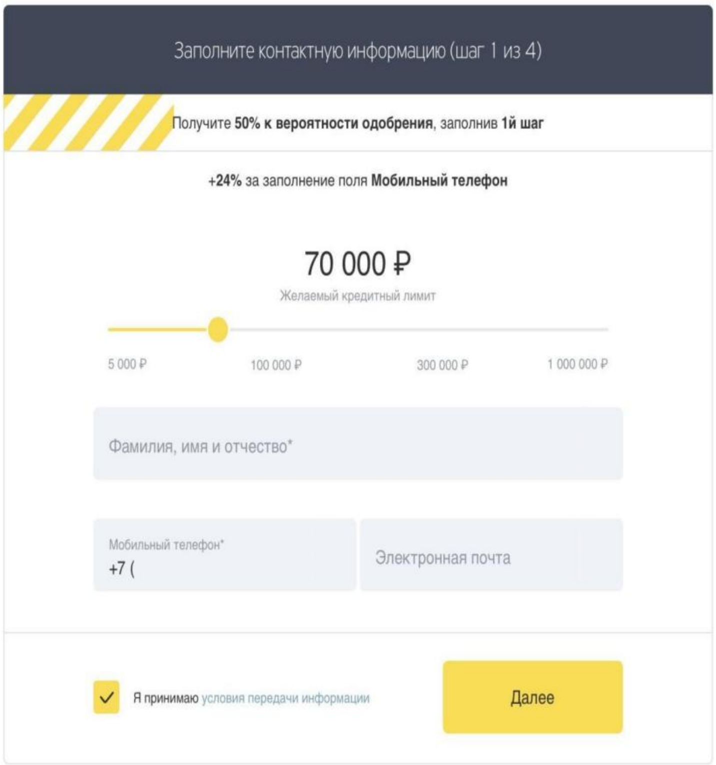
Introduction

In business, numerous ideas constantly emerge. The main purpose of these ideas is to improve the company's key performance indicators. The primary goal of a company that tests these ideas is to implement only successful concepts. In the modern world of business and internet marketing, decisions are made based on scientific data and research on user behavior. A/B testing is a discipline that allows companies to obtain this scientific data. A/B testing is an experiment conducted to determine which of two (or more) variants of a web page, application, or advertisement improves the target metric.

The process can be briefly described as follows:

1. Splitting a sample of users into groups, usually into a control and a test group.
2. Demonstration of one of the variations to each group.
3. Getting the result and interpreting it.

A web form is a part of a website or application that allows users to enter and send data to a server. A web form usually consists of several consecutive pages. It can contain text input fields, radio buttons for selecting options, buttons for uploading files, and other controls. Web forms are often used in online stores when placing orders. In view of this, companies try to make them as convenient and simple as possible so that customers do not leave at the stage of filling out the application. This is very important because filling out a web form is often perceived as something routine. Any difficulties and inconveniences will alienate potential customers. A/B testing is used to create and optimize such a form. An example of what a web form looks like is shown in Figure 1.



Заполните контактную информацию (шаг 1 из 4)

Получите **50% к вероятности одобрения**, заполнив **1й шаг**

+24% за заполнение поля **Мобильный телефон**

70 000 ₽

Желаемый кредитный лимит

5 000 ₽ 100 000 ₽ 300 000 ₽ 1 000 000 ₽

Фамилия, имя и отчество*

Мобильный телефон*
+7 (

Электронная почта

☑ Я принимаю условия передачи информации

Рисунок 1 - пример веб-формы

A lot of metrics are being tested for web forms, for example:

1. The average duration of filling out the form.
2. The average duration of filling out a certain page of a web form.
3. Conversion to form filling.
4. Conversion to filling out a specific page of a web form.
5. Conversion to purchase (that is, the proportion of forms that completed the purchase from all forms).
6. The frequency of occurrence of any alert on the form or page.

The relevance of the work

A/B testing provides an opportunity for developers and marketers to find the most effective methods of attracting traffic to the site and increasing the number of customers, as well as increasing user satisfaction and improving the user experience. Nowadays, A/B testing has become a mandatory stage in the creation and optimization of websites, applications, e-shops, social networks and many other online platforms. This test allows you to get rid of guesses when choosing the most suitable idea and gives a clear answer, mathematically justified. Now in large companies there are entire analytical teams that do this, which confirms the relevance of the method. The most popular examples of A/B testing can be identified:

1. A/B Price testing - Stores use A/B testing to determine the most effective price for a product. To do this, several different product prices are tested to see which price is best accepted by customers.
2. A/B Image Testing - Marketing companies conduct A/B testing to understand which image attracts more customers.
3. A/B Headline testing - Companies test the headlines of ads and articles to determine which headline attracts attention better.
4. A/B testing of web forms - companies test web forms in order to explore key metrics both on each individual page and on the entire form as a whole. The importance of their improvements has been described before.

This topic has a lot of scientific research. A striking example is the work of Ron Kohavi “Controlled experiments on the web: survey and practical guide. Data Mining and Knowledge Discovery”. It describes a practical guide for conducting controlled experiments on the Internet. The authors review the methodologies for conducting A/B testing on websites and applications, describe in detail the basic concepts in this area and provide numerous recommendations and tools for successful A/B testing. The article also contains illustrative examples of research by large global companies, thus confirming the relevance of the method.

Goals and objectives of the work

The purpose of this work is to explore the possibilities of modern A/B testing approaches and their application to web forms.

To achieve the above goal, the following tasks were set:

- Analysis of current researches in the field of A/B testing.
- Analysis of user data of web forms that are planned to be investigated as part of this work.
- Coding of the studied methods.
- Testing the collected data.
- Summarizing the results.

Chapter 1. Analysis of modern A/B testing methods

1.1 Process of testing

A/B testing is a reliable and effective tool for optimizing brand interaction with consumers in a digital environment. By conducting experiments and testing business hypotheses, we optimize various points of contact with consumers, such as a website, an application, advertising banners, texts of advertising messages and others. In this paper, we will consider A/B testing using two variations.

The A/B test begins by identifying the problem that we want to solve with it. An example could be as follows. We have a web form that consists of several pages. The first page lists mobile Internet rates, as well as fields for entering personal data. The following pages also contain fields for entering some data, and the last one has a button, when clicked, the form is considered completed. We will call the completed form a full application. The problem is that the conversion from a visit (opening a web form) to a full application is much lower than on other forms on our site.

Then hypotheses are formed. The hypothesis implies a solution to our problem. Companies often use certain templates to form it. For example: "if we do something, something will happen, and therefore it will affect our metrics in this way." Continuing the example with the web form, it may look like this. If we add navigation at the top of the form, it will become more convenient for the client to understand how many stages are left until the end of the form, and the conversion to a full application will increase.

To form hypotheses, they resort to the following things: data analysis, conducting surveys, UX research, CustDev research, market and competitor analysis, analysis of reviews and requests to technical support, and so on. Obviously, this requires not only analysts, but also other teams.

The next step is to think over possible options for the failure of the chosen solution. For example, the page loading time may increase. It is not always possible to predict all the options in advance, but at least some should be highlighted in order to simplify your work in the future.

Next, we define metrics, they are divided into basic and additional ones. The main one is the one by which we will determine which of the options won. Additional metrics allow you to verify that the test has not degraded other processes. In the example with the web form, the main metric is the conversion to a full application. Conversion to utilizing can be considered as an additional option. Conversion to utilizing is the ratio of the number of "utilized users" to the total number of users. The concept of a utilized user may vary in depending on the situation. In our case, this is the person who filled out and submitted the form, but did not use the product further, and the total number of users is all the people who filled out the web form. This metric allows you to check whether a new effect has attracted a disinterested audience.

There are a huge number of metrics that are used in A/B testing. It all depends on the specifics of the product. The main ones can be distinguished:

1. ARPU, Average Revenue Per User - the average income per customer for a certain period of time.
2. ARPU, Average Revenue Per Paying User - the average income per paying customer for a certain period of time.
3. Various conversions
4. Duration of the session

The next step is to prioritize the test. A huge number of A/B tests are conducted in large companies. Several A/B tests may be planned for one product, but they cannot be performed simultaneously on the same users, because in this case we will not understand which hypothesis influenced the result. In this regard, we prioritize the tests in order to understand in which order to run them. An example is the RICE methodology. In it, the formula (1) is calculated for each test. The test with the highest value is started first. Reach - the number of users we want to influence. Impact is an assessment of how much our change will affect user behavior. The score can be equal to one of the following values: 3 – "mass impact", 2 – "high", 1 – "average", 0.5 – "low", 0.25 – "minimum". Confidence – an assessment of how confident we are in our decision. This estimate has a value in the range [0.01, 10]. Effort shows how many people/hours it will take to prepare the change.

$$\frac{REACH \times IMPACT \times CONFIDENCE}{EFFORT} \tag{1}$$

Next, you need to determine the statistical criterion and the duration of the test. The duration of the test means the number of users who will participate in the test. To determine a statistical criterion, you need to do some analysis on historical data. You need to understand what kind of character they have. To do this, simulations of A/A and A/B testing are carried out. A description of these processes will be given in the practical part. The duration of the test is calculated only after determining the statistical criterion. To calculate it, you need to determine the minimum detectable effect and some other parameters that depend on the selected metric. The minimum detectable effect in A/B testing is the lowest value that the company expects to receive from changes made to the tested version of the site or application. This indicator is usually set in advance and determined based on historical data, business goals and expectations. Thus, two hypotheses are put forward. The null hypothesis of the test states that there is no statistically significant difference between the control and test variations, and the alternative hypothesis says that there is a difference. A/B testing uses a variety of statistical criteria, the main ones are:

1. Z-criterion for two proportions
2. Student's T-test
3. The Mann-Whitney U-test
4. Permutation test
5. Bootstrap test

In addition, other criteria of mathematical statistics are used. For example, the Shapiro-Wilk test for the normality of a distribution, the Fisher test for the equality of variances of two normal distributions, and others.

Next, the required amount of data is collected and statistics are calculated. The result must be interpreted, and p-value is used for this. The P-value (significance level) is the probability of getting a result that is more extreme than the one that was obtained, provided that the null hypothesis is true. In the context of A/B testing, p-value is used to determine the statistical significance of the differences between the control and test groups. If the p-value is below the selected significance level (usually equal to 0.05), then it is considered that the differences between the groups are statistically significant (there is a significant difference between the two groups). If the p-value is above the significance level, then the null hypothesis cannot be rejected and the differences between the groups are not statistically significant.

1.2 Modern approaches in A/B testing

1.2.1 Variance reduction

In A/B testing, great attention is paid to reducing the variance of the metric. The smaller the variance, the more sensitive our test will be. It also reduces the required number of observations in the samples. There are various approaches to reduce variance, the most relevant being stratification and CUPED. It is important to note that these techniques preserve the unbiased estimates. The articles devoted to these methods show that in the worst case, the variance will not change, but in most cases it will decrease. The effect on sensitivity is extremely positive.

The essence of stratification is that we assume that users can be divided into groups(strata) according to some criteria. For example, by region. If we assume that there is a difference in the metric between such groups, then stratification greatly reduces the variance. In this method, for test and experimental samples, we take representatives of each stratum in an amount proportional to their fractions in the general population. This is called stratified sampling. Next, we calculate for each sample not the usual average, but a stratified one, which in turn is simply weighted, where the weight of the stratum is the proportion of the stratum in the population. The population is all the users that we can influence with our experiment. For example, the Student's criterion can already be applied to the obtained metrics. The estimate of the stratified mean is unbiased, the estimate of the variance of this estimate is known. The best thing to find is Huizhi Xie's article “Increasing the Sensitivity of Controlled Online Experiments: Case Studies on Netflix.”

In CUPED, to reduce the variance, we try to get rid of the effect of pre-experimental data. We don't want them to affect the current test in any way. We do this with the help of some mathematical transformations based on the properties of variance and covariance. To do this, for each user participating in an A/B experiment, you need to know this metric in the pre-experimental period and in the current one. Y is the user's metric for the current experiment, X is the user's metric for the pre-experimental period. Formula (2) shows the necessary transformation to obtain a new metric. The variance of the new metric is derived in formula (3). It can be seen that this variance is less than or equal to the variance of the original metric. After the transformations, the Student's criterion can be applied. A more detailed description of the approach is described in the articles by Alex Deng “Improving the Sensitivity of Online Controlled Experiments: Case Studies at Netflix” and Simon Jackson “How Booking.com increases the power of online experiments with CUPPED".

$$Y_{new} = Y - \theta X \tag{2}$$

$$\theta = \frac{cov(Y, X)}{var(X)}$$

$$var(Y_{new}) = var(Y) * (1 - corr^2(Y, X)) \tag{3}$$

1.2.2 Dealing with ratio-metrics

A ratio metric is a metric of a non-user-level ratio with dependent observations, but which is explicitly expressed through the ratio of the sums of the corresponding user signals. The problem with such a metric is that we cannot estimate its variance of the mean in the usual way. There is a special delta method that makes this estimate. Formula (4) shows this estimate of variance. Using these estimates, a proportional z-test can be performed. The method is described in more detail in the articles by Roman Budylin "Consistent Transformation of Ratio Metrics for Efficient Online Controlled Experiments" and Alex Deng "Applying the Delta Method in Metric Analytics: A Practical Guide with Novel Ideas".

$$var(\frac{X}{Y}) \approx \frac{1}{E[Y]^2} Var(X) + \frac{E[X]^2}{E[Y]^4} var(Y) - 2 * \frac{E[X]}{E[Y]^3} cov(X, Y) \tag{4}$$

The next approach is called linearization. It converts a ratio metric with dependent observations into an average user metric with independent ones. Formula (5) shows this transformation, where the letter u denotes a unique user. The difference in linearized metrics always remains aligned with the change in the target ratio metric. Linearized user signals can already be considered independent and their statistical significance can be determined by the Student's test. At the same time, the p-value values for the linearized metric will be consistent with the values obtained using the delta method on the original ratio metric.

$$\frac{\sum\limits_u X(u)}{\sum\limits_u Y(u)} \approx avg_u(L(X(u), Y(u))) \tag{5}$$

$$L(X(u), Y(u)) \approx X(u) - kY(u)$$

$$k = \frac{\sum\limits_{u} X(u)}{\sum\limits_{u} Y(u)} \qquad (6)$$

It is noteworthy that in formula (6) we use data only from the control sample. It can also be noted that the value of k can be calculated using a regression model in which X(u) is an independent variable, Y(u) is a dependent variable, and k is the coefficient before the independent variable.

Another approach is to bootstrap ratio metrics. This is the same A/B test using the statistical bootstrap method, only at each step we now calculate the ratio metrics for the control and test samples. It is important to note that now, for each step of the cycle, we sample samples with repetitions by objects, and not by observations. If the object is in a test or control group, then all its actions will be in this group.

Keyu Nie's article “Dealing With Ratio Metrics in A/B Testing at the Presence of Intra-User Correlation and Segments” describes an algorithm for working with ratio metrics and suggests their unbiased estimation of mathematical expectation. The main difference from the previous methods is that the article proves that this estimate has the lowest variance, and so that, the estimate is effective. Their estimates of variances are also given in this article.

1.2.3 Applying of multi-armed bandits method

The essence of the method is that we dynamically determine which group to send the current user to during the A/B test. This is done because we do not want to send a large amount of traffic to obviously bad options. This method has received the greatest demand in conversion tests. The bottom line is as follows. It is convenient to approximate the conversion distribution using beta distribution. At the very beginning, for each test variant, we introduce two parameters of the beta distribution and equate them to one (7). With such parameters, the beta distribution becomes normally uniform. We set such parameters because initially we don't know how the conversion of each option behaves. Next, we repeat the following steps for each new user. First, for each test's variant, we sample the conversion from the beta distribution with the parameters that are currently relevant (8). Next, we select the variant with the highest conversion and show it to the current user. Next, we get either y=1, which means that the user has “passed” the conversion, or y=0, which means otherwise. We update the parameters of this test variant according to the formula (9). When we have run out the number of users we can afford to run the A/B test, the test ends. The option with the highest conversion rate at the moment wins. There is no “statistically significant difference” option here.

$$\forall i, \alpha_i = \beta_i = 1 \tag{7}$$

$$\theta_i \sim Beta(\alpha_i, \beta_i) \tag{8}$$

$$k = argmax(\theta_i)$$

$$\alpha_i = \alpha_i + y; \quad \beta_i = \beta_i + 1 - y \tag{9}$$

It is proved that in the case of the beta distribution, as described above, with the number of users in the A/B test tending to infinity, the probability that the algorithm will come to the correct solution is one.

There are also modifications to the method that allow us to monitor the rest of the groups, and if we were unlucky at first, or the trend for choosing an option changed, we could feel it and fix the situation. A full description of the algorithm is contained in Maryam Aziz's article “On Multi-Armed Bandit Designs for Dose-Finding Clinical Trials.

The main disadvantage of this method is the inability to obtain an accurate value of the number of users necessary to fix a statistically significant difference with a given minimum detectable effect, significance level and power.

1.2.4 Peeking problem

If we have the opportunity to observe the dynamics of p-value changes during the A/B test, then there may be a desire to prematurely stop the experiment whenever we see a statistically significant result. But that would be a mistake. The more often we look at the intermediate A/B test results with a willingness to make a decision based on them, the more likely it becomes that the criterion will show a significant difference when there is none. It is theoretically proven that if you run an A/A test (that is, show both groups the same version of the page), then the p-value will cross the threshold at least once with the number of users tending to infinity. However, there are methods that allow you not to wait until the end of the test.

Peter O'Brien's article “A Multiple Testing Procedure for Clinical Trials” describes the Pocock and O'Brien-Fleming methods. For both amendments, you need to know in advance the maximum duration of the test and the number of p-value checks between the start and end of the test. Moreover, the checks should take place through equal numbers of observations. The article by Wenru Zhou “Approaches to analyzing binary data for large-scale A/B testing" shows that the corrections work correctly. In the Pocock method, we summarize the results of the tests through equal numbers of observations, but with some reduced (stricter) significance level. In the O'Brien-Fleming method, the significance level varies depending on the verification number. The earlier we try to complete the test, the more stringent the threshold will be applied.

Ramesh Johari's article “Peeking at A/B Tests” describes the SPRT method and its modification mSPRT. This is a sequential test that uses the maximum likelihood function in its algorithm. This amendment is the most relevant at the moment in the industry.

1.2.5 Analysis of P-values based on historical A/B tests

When a large number of tests have already been conducted in the company, you can assess the overall situation and understand how well they pass.

Formula (10) shows what, by definition, we should have a p-value distribution on a specific test, where the hypothesis H0 is correct. It can be seen that this is a function of uniform distribution. So that, if we take all the p-values in the tests where we did not get a statistically significant result, we should see their uniform distribution.

$$P(pvalue < \alpha | H_0) = \alpha \quad \forall \alpha \in [0, 1] \tag{10}$$

$$P(pvalue \geq A | H_1) = B \Rightarrow P(pvalue < A | H_1) = 1 - B \tag{11}$$

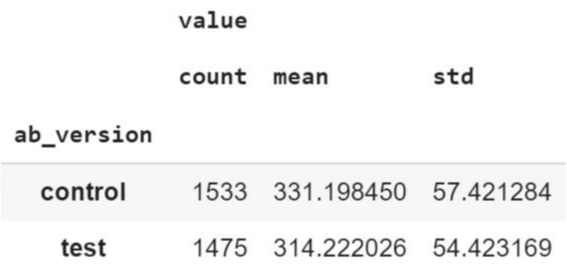
On the other hand, formula (11) shows the distribution of p-value, provided that the hypothesis H1 is correct. A is the probability of a type I error, B is the probability of a type II error. We can conclude that if we take all the p-values in the tests where we got a statistically significant result, then these p-values should “shrink” to zero.

Chapter 2. Program realization of A/B testing

2.1 A/B testing of form filling time

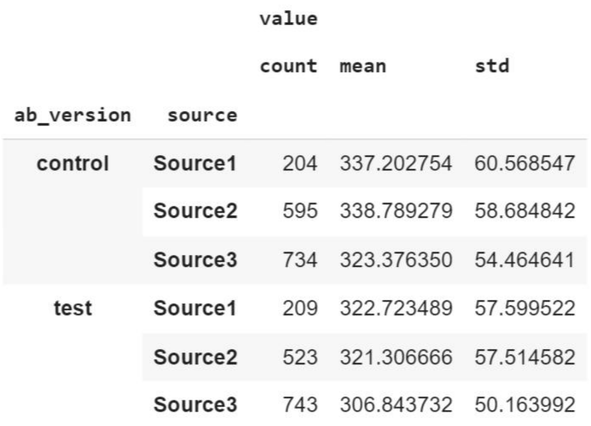
For this test we use a dataset which contains data on how users behave in the process of filling out a web form of some service. They fill out this web form every month, it provides updating of some user data. Some improvements have been proposed, which are suggested to be tested in the A/B test. The dataset consists of the following columns: device - which device the user uses when filling out a web form, there may be 2 options – “desktop” or “mobile"; source - the source from which the user came to the service initially; ab_version - can be either “test” (in this case, the version of the web form with changes is displayed to users) or “control” (in this case, the version of the web form is displayed to users without changes); value - the amount of time in seconds it took the user to fill out the web form; value_before - the average amount the time in seconds for the previous 3 months that the user needed to fill out the web form. The hypothesis is that the new changes will reduce the average time to fill out a web form. The entire implementation code is written in the Python programming language.

Figure 2 shows the breakdown by test and control with the number of users, the average time value and the standard deviation of the time. In Figures 3 and 4, the source and device breakdowns are additionally added, respectively. You can notice the dependence of the average time on both the source from which the users came and the device. This suggests that a stratified Student's test can give better results on a given dataset.



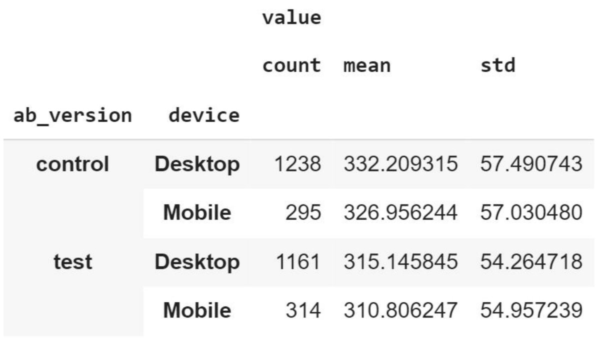
	value		
	count	mean	std
ab_version			
control	1533	331.198450	57.421284
test	1475	314.222026	54.423169

Рисунок 2 - разбивка на тест и контроль



		value		
		count	mean	std
ab_version	source			
control	Source1	204	337.202754	60.568547
	Source2	595	338.789279	58.684842
	Source3	734	323.376350	54.464641
test	Source1	209	322.723489	57.599522
	Source2	523	321.306666	57.514582
	Source3	743	306.843732	50.163992

Рисунок 3 - дополнительная разбивка по источнику



		value		
		count	mean	std
ab_version	device			
control	Desktop	1238	332.209315	57.490743
	Mobile	295	326.956244	57.030480
test	Desktop	1161	315.145845	54.264718
	Mobile	314	310.806247	54.957239

Рисунок 4 - дополнительная разбивка по устройству

Figure 5 shows the value_before distributions for the test and control. They look quite similar. This is obvious, since users had the same versions of the web form during these periods. The opposite can be seen in Figure 6, which shows the value distributions for the test (red) and control (blue). But only an A/B test can show whether the result is statistically significant.

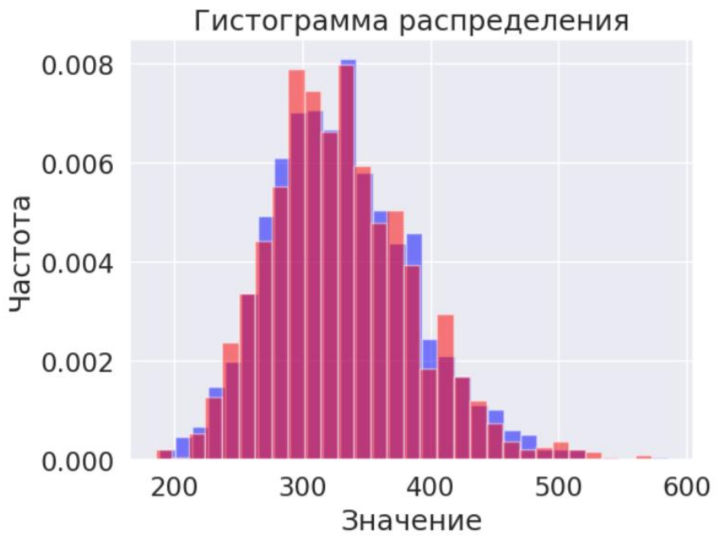


Рисунок 5 - распределения value_before

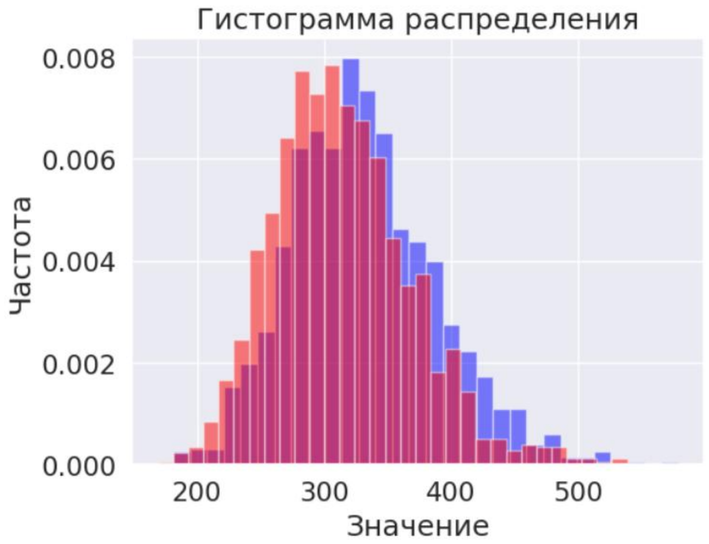


Рисунок 6 - распределения value

Let's check the test and control for the normality of the distribution and equality of variances. The qq-plot shown in Figures 7 and 8 was used to check for normality. Both graphs show that the distributions are normal. To check for equality of variances, we used the Fisher test on historical data (value_before) with the null hypothesis that the variances are equal, we got a p-value equal to 0.31, which is greater than 0.05. Accordingly, the variances are equal. This gives us the right to use the usual implementation of the Student's test with the assumption of normal distributions and equal variances.

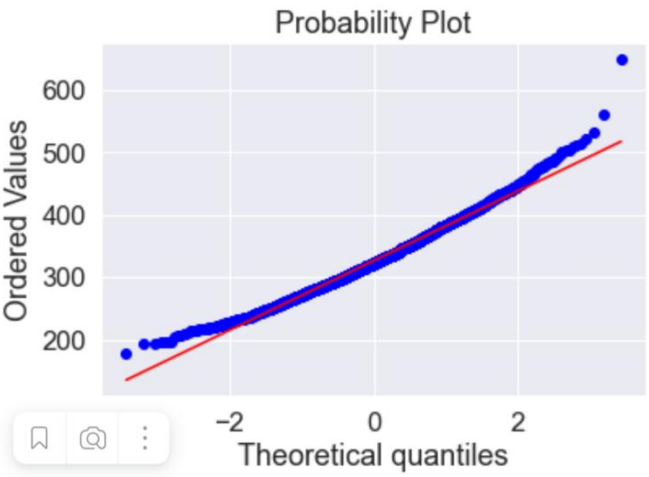


Рисунок 7 - qq-plot для теста

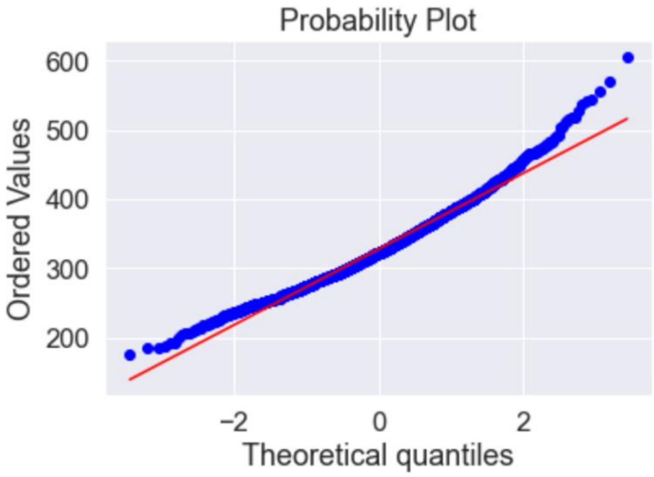


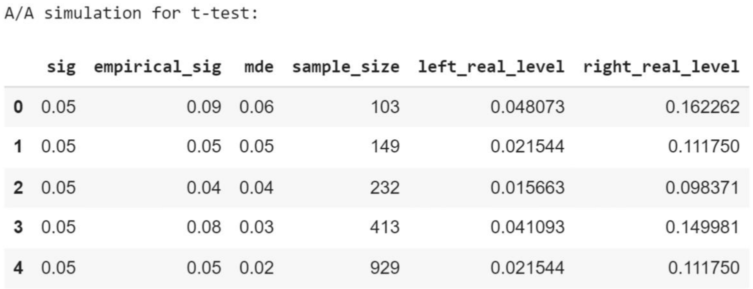
Рисунок 8 - qq-plot для контроля

The following simulation of the A/B test was implemented for all statistical criteria except msprt. Several minimal detectable effects (mde) are iterated through in the loop, which should be considered statistically significant. In this A/B test it is: 0.06, 0.05, 0.04, 0.03, 0.02. At each such iteration of the cycle, an artificial test sample is created, which consists of control values, from which the average control value multiplied by mde was subtracted, and the control itself remains the same. Next, the minimum sample size required to detect a given effect is calculated. An error of the I type is equated to 0.05, an error of the II type to 0.2. After that, a cycle is created and samples with repetitions are taken N times for control and artificial test, which are used to calculate the statistics of a certain criterion, then p-value is calculated at each of the N steps, which show whether the results are statistically significant. All N results are averaged and an empirical power is obtained, which, if the test is correct, should be equal to the theoretical one.

The difference between the simulation of an A/A test and the simulation of an A/B test is only that the artificial test here simply equals the control. The A/A test checks whether the empirical I type error is equal to the theoretical one. Separate simulations are used for msprt, since this is a sequential criterion and it is impossible to determine the sample size there in advance. Also, in the simulation of the A/A test for msprt, various MDEs are not iterated over. This is not necessary, since we do not calculate the sample size in advance.

Next, for all statistical criterions, the results of A/A and A/B test simulations will be shown in the form of tables. Successful options are those in which the empirical levels of significance and power are equal to the theoretical ones. The field designations in the table are: sig - theoretical I type error, empirical_sig - empirical I type error obtained in simulation of A/A test; empirical_power - empirical power obtained in simulation of A/B test; left_real_level and right_real_level - respectively, the left and right boundaries of the confidence interval are either empirical_sig or empirical_power; mde is the minimum detectable effect; sample_size is the sample size that is used for this mde.

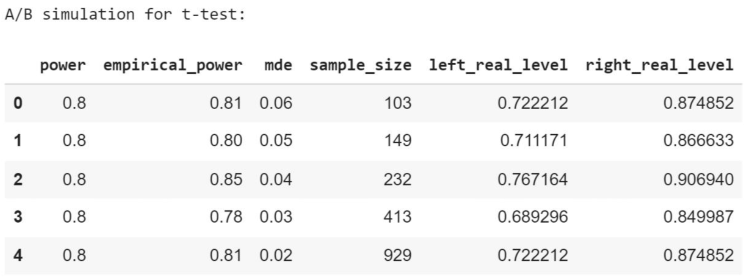
Figures 9 and 10 show the results for the Student's test. The empirical I type error is very close to the theoretical one, while the empirical power shows higher results than the theoretical one.



A/A simulation for t-test:

	sig	empirical_sig	mde	sample_size	left_real_level	right_real_level
0	0.05	0.09	0.06	103	0.048073	0.162262
1	0.05	0.05	0.05	149	0.021544	0.111750
2	0.05	0.04	0.04	232	0.015663	0.098371
3	0.05	0.08	0.03	413	0.041093	0.149981
4	0.05	0.05	0.02	929	0.021544	0.111750

Рисунок 9 - A/A симуляция для теста Стьюдента

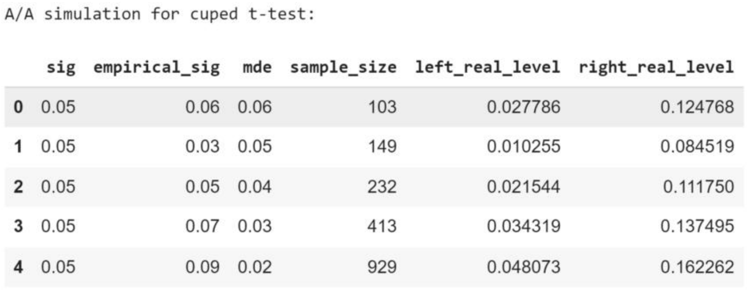


A/B simulation for t-test:

	power	empirical_power	mde	sample_size	left_real_level	right_real_level
0	0.8	0.81	0.06	103	0.722212	0.874852
1	0.8	0.80	0.05	149	0.711171	0.866633
2	0.8	0.85	0.04	232	0.767164	0.906940
3	0.8	0.78	0.03	413	0.689296	0.849987
4	0.8	0.81	0.02	929	0.722212	0.874852

Рисунок 10 - A/B симуляция для теста Стьюдента

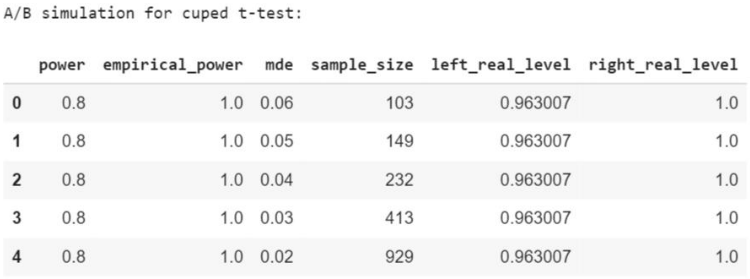
Figures 11 and 12 show the results for the Student's CUPED test. As mentioned earlier, we have averaged data for each user in the periods preceding the A/B test. This can be used in the CUPED method to increase empirical power. The empirical I type error turned out to be approximately equal to the theoretical one, while the empirical power shows a result equal to one, which indicates that the test does not make II type errors.



A/A simulation for cuped t-test:

	sig	empirical_sig	mde	sample_size	left_real_level	right_real_level
0	0.05	0.06	0.06	103	0.027786	0.124768
1	0.05	0.03	0.05	149	0.010255	0.084519
2	0.05	0.05	0.04	232	0.021544	0.111750
3	0.05	0.07	0.03	413	0.034319	0.137495
4	0.05	0.09	0.02	929	0.048073	0.162262

Рисунок 11 - A/A симуляция для теста CUPED

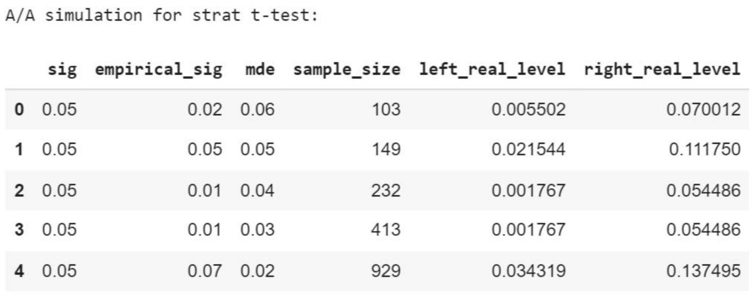


A/B simulation for cuped t-test:

	power	empirical_power	mde	sample_size	left_real_level	right_real_level
0	0.8	1.0	0.06	103	0.963007	1.0
1	0.8	1.0	0.05	149	0.963007	1.0
2	0.8	1.0	0.04	232	0.963007	1.0
3	0.8	1.0	0.03	413	0.963007	1.0
4	0.8	1.0	0.02	929	0.963007	1.0

Рисунок 12 - A/B симуляция для теста CUPED

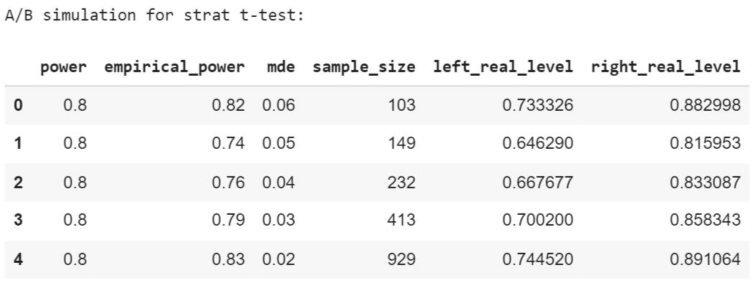
It was previously noted that there is some dependence in the data on the time of passing the web form from the type of device and source. This suggests that a stratified Student test can give better results than a regular one. Figures 13 and 14 show the results of a test in which the device is used as a stratum. There are noticeable improvements in the empirical I type error. Figures 15 and 16 show the results of a test in which the source is used as a stratum. There are noticeable improvements in both the empirical I type error and the empirical power.



A/A simulation for strat t-test:

	sig	empirical_sig	mde	sample_size	left_real_level	right_real_level
0	0.05	0.02	0.06	103	0.005502	0.070012
1	0.05	0.05	0.05	149	0.021544	0.111750
2	0.05	0.01	0.04	232	0.001767	0.054486
3	0.05	0.01	0.03	413	0.001767	0.054486
4	0.05	0.07	0.02	929	0.034319	0.137495

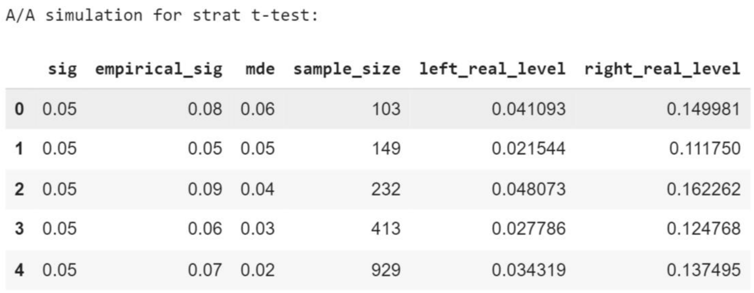
Рисунок 13 - A/A симуляция для стратифицированного теста Стьюдента, где устройство является стратой



A/B simulation for strat t-test:

	power	empirical_power	mde	sample_size	left_real_level	right_real_level
0	0.8	0.82	0.06	103	0.733326	0.882998
1	0.8	0.74	0.05	149	0.646290	0.815953
2	0.8	0.76	0.04	232	0.667677	0.833087
3	0.8	0.79	0.03	413	0.700200	0.858343
4	0.8	0.83	0.02	929	0.744520	0.891064

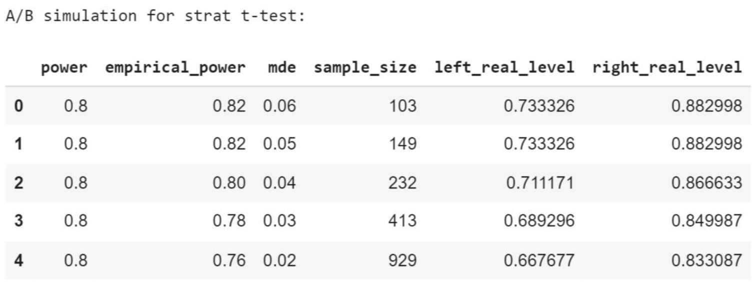
Рисунок 14 - A/B симуляция для стратифицированного теста Стьюдента, где устройство является стратой



A/A simulation for strat t-test:

	sig	empirical_sig	mde	sample_size	left_real_level	right_real_level
0	0.05	0.08	0.06	103	0.041093	0.149981
1	0.05	0.05	0.05	149	0.021544	0.111750
2	0.05	0.09	0.04	232	0.048073	0.162262
3	0.05	0.06	0.03	413	0.027786	0.124768
4	0.05	0.07	0.02	929	0.034319	0.137495

Рисунок 15 - A/A симуляция для стратифицированного теста Стьюдента, где источник является стратой

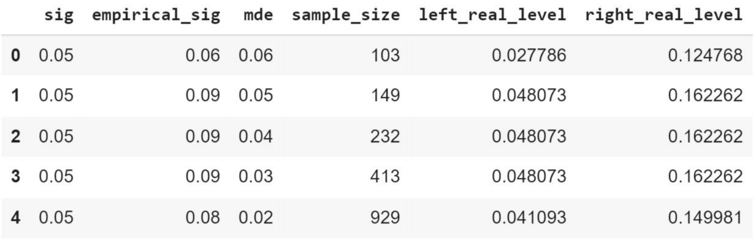


A/B simulation for strat t-test:

	power	empirical_power	mde	sample_size	left_real_level	right_real_level
0	0.8	0.82	0.06	103	0.733326	0.882998
1	0.8	0.82	0.05	149	0.733326	0.882998
2	0.8	0.80	0.04	232	0.711171	0.866633
3	0.8	0.78	0.03	413	0.689296	0.849987
4	0.8	0.76	0.02	929	0.667677	0.833087

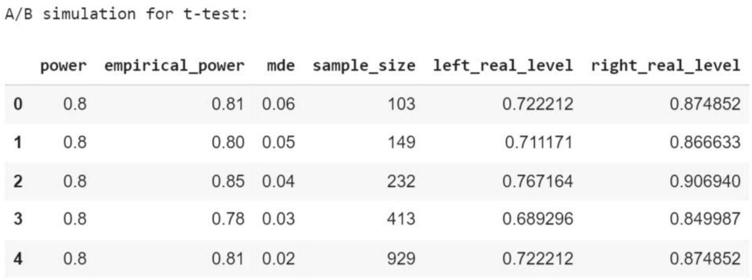
Рисунок 16 - A/B симуляция для стратифицированного теста Стьюдента, где источник является стратой

Figures 17 and 18 show the results for the Bootstrap test. It is usually used in situations where we cannot use a regular Student's test. For example, if our data does not follow a normal distribution or the variance of the test and control are not equal. Therefore, it is obvious that the empirical results are also close to the theoretical ones, but there are no noticeable improvements. It can also be noted that a sample size calculation formula was used here for the usual Student's test. Since the simulations have shown satisfactory results, it is possible to use this formula with this dataset.



	sig	empirical_sig	mde	sample_size	left_real_level	right_real_level
0	0.05	0.06	0.06	103	0.027786	0.124768
1	0.05	0.09	0.05	149	0.048073	0.162262
2	0.05	0.09	0.04	232	0.048073	0.162262
3	0.05	0.09	0.03	413	0.048073	0.162262
4	0.05	0.08	0.02	929	0.041093	0.149981

Рисунок 17 - А/А симуляция для теста Бутстрап

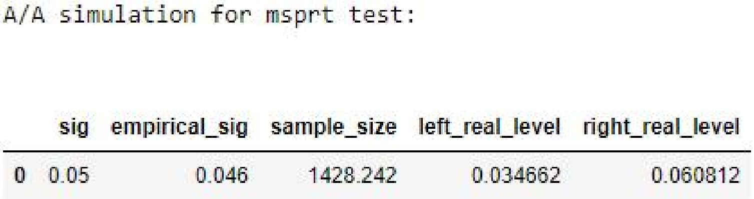


A/B simulation for t-test:

	power	empirical_power	mde	sample_size	left_real_level	right_real_level
0	0.8	0.81	0.06	103	0.722212	0.874852
1	0.8	0.80	0.05	149	0.711171	0.866633
2	0.8	0.85	0.04	232	0.767164	0.906940
3	0.8	0.78	0.03	413	0.689296	0.849987
4	0.8	0.81	0.02	929	0.722212	0.874852

Рисунок 18 - А/В симуляция для теста Бутстрап

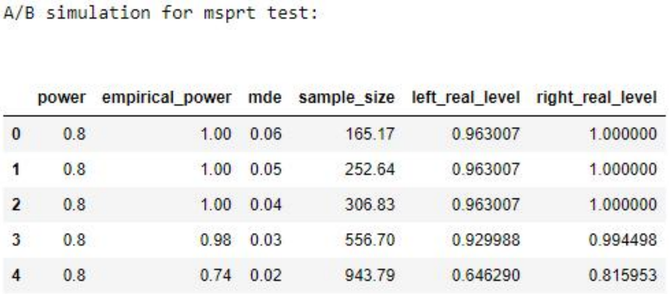
The last test is msprt. Figures 19 and 20 show the results. The sample_size field here indicates the average sample size of all samples that were used in N iterations of the simulation. It can be seen that the empirical error of the first kind is equal to the theoretical one, and the power is close to one. Sample sizes also turn out to be acceptable. However, this method has a disadvantage - it is impossible to determine in advance the sample size necessary to fix a particular mde.



A/A simulation for msprt test:

	sig	empirical_sig	sample_size	left_real_level	right_real_level
0	0.05	0.046	1428.242	0.034662	0.060812

Рисунок 19 - A/A симуляция для теста msprt



A/B simulation for msprt test:

	power	empirical_power	mde	sample_size	left_real_level	right_real_level
0	0.8	1.00	0.06	165.17	0.963007	1.000000
1	0.8	1.00	0.05	252.64	0.963007	1.000000
2	0.8	1.00	0.04	306.83	0.963007	1.000000
3	0.8	0.98	0.03	556.70	0.929988	0.994498
4	0.8	0.74	0.02	943.79	0.646290	0.815953

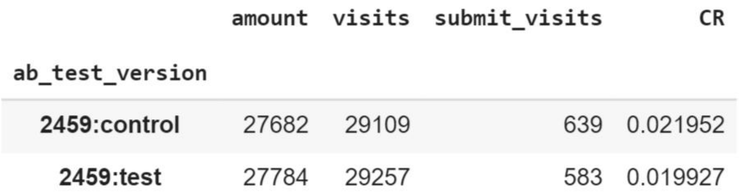
Рисунок 20 - A/B симуляция для теста msprt

Based on the results of the simulations, it was decided to use the CUBED criterion for the A/B test. It shows the best results in A/A and A/B simulations, and also requires acceptable sample sizes. In our case, we will consider the minimum detectable effect to be 0.03. The test showed a statistically significant difference, p-value = 0.01 (less than 0.05). For this, it required 413 elements of each sample.

2.2 A/B testing the conversion from a web form visit to an application submission

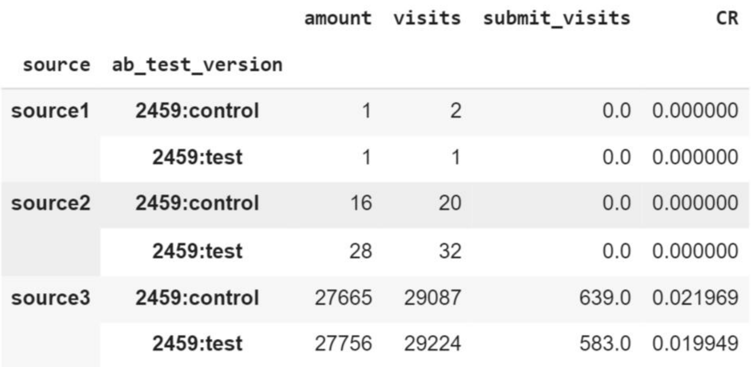
For this test, a dataset is used, which also contains data on how users behave in the process of filling out a web form for some service. The dataset consists of the following columns: hit-dttm – date and time of the user's action, source – the source from which the user came; ab_test_version – version of A/B test (we are interested in the values “2459:control” and “2459:test”); device_type – which device the user uses when filling out the web form, there may be 2 options - “desktop” or “mobile”; visitor_id – the user's unique number; visit_num is the number of the user's session during which this action was performed; event_name is the type of action performed by the user (one of the following options: “pageview", “formStep2", “formStep3", “formStep4", “formSubmit"). The metric that will be tested is the conversion from a web form visit to an application submission, which means dividing the total number of applications by the total number of visits. The “pageview” action corresponds to the visit of the web form, and the “formSubmit” action corresponds to the submission of the application. It is important to note that one user can open a web form several times, but in different sessions. Accordingly, we are dealing with a ratio metric.

Figure 21 shows the breakdown by test and control with the number of users, the number of visits, the number of applications sent, and conversions. In Figures 22 and 23, the source and device breakdowns are additionally added, respectively. You can notice the dependence of the conversion on the device.



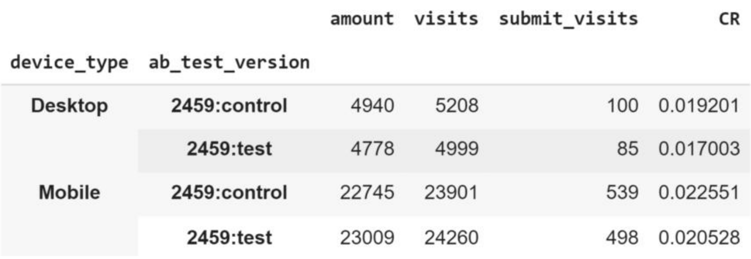
	amount	visits	submit_visits	CR
ab_test_version				
2459:control	27682	29109	639	0.021952
2459:test	27784	29257	583	0.019927

Рисунок 21 - разбивка на тест и контроль



		amount	visits	submit_visits	CR
source	**ab_test_version**				
source1	**2459:control**	1	2	0.0	0.000000
	2459:test	1	1	0.0	0.000000
source2	**2459:control**	16	20	0.0	0.000000
	2459:test	28	32	0.0	0.000000
source3	**2459:control**	27665	29087	639.0	0.021969
	2459:test	27756	29224	583.0	0.019949

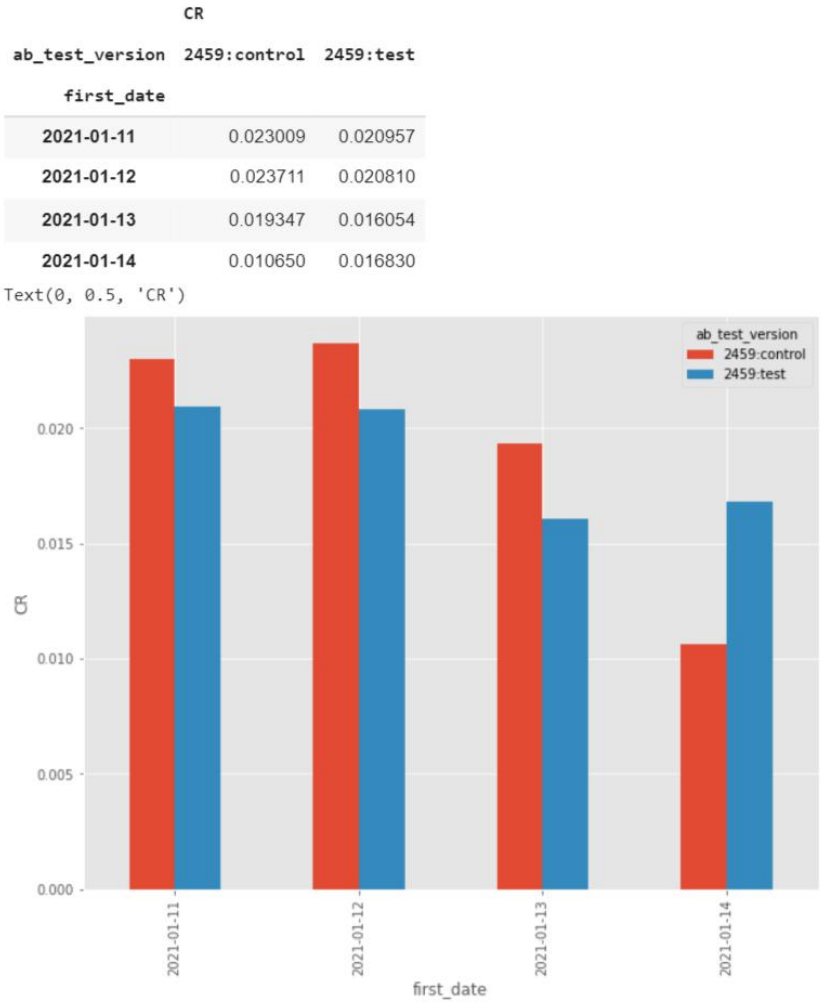
Рисунок 22 - дополнительная разбивка по источнику



		amount	visits	submit_visits	CR
device_type	**ab_test_version**				
Desktop	**2459:control**	4940	5208	100	0.019201
	2459:test	4778	4999	85	0.017003
Mobile	**2459:control**	22745	23901	539	0.022551
	2459:test	23009	24260	498	0.020528

Рисунок 23 - дополнительная разбивка по устройству

Figure 24 shows conversions by date. It is noteworthy that on the last day of testing, user behavior was different from user behavior in previous days. This is probably due to the relatively small number of users: 1784 in the control sample and 1842 in the test sample.

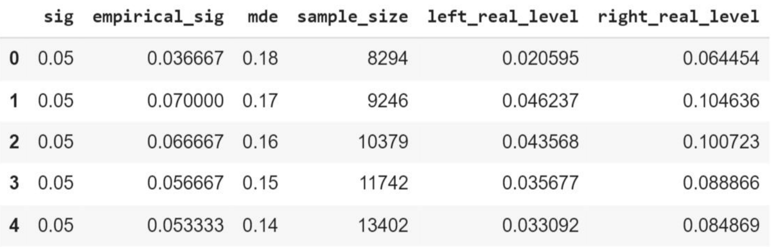


	CR	
ab_test_version	2459:control	2459:test
first_date		
2021-01-11	0.023009	0.020957
2021-01-12	0.023711	0.020810
2021-01-13	0.019347	0.016054
2021-01-14	0.010650	0.016830

```
Text(0, 0.5, 'CR')
```

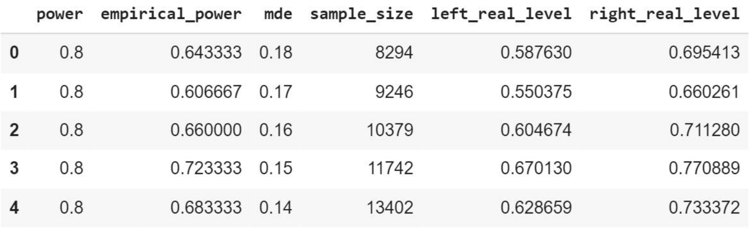
Рисунок 24 - конверсия в разбивке по датам

Simulations of A/A and A/B tests will occur according to a similar algorithm described in Chapter 2.1. First, consider the proportional z-test, its simulations are shown in Figures 25 and 26, respectively. It can be seen that the empirical power is on average less than the theoretical power (which is 0.8). This is probably due to the fact that we have a ratio metric. Theoretically, it is considered incorrect to use this test for such a metric.



	sig	empirical_sig	mde	sample_size	left_real_level	right_real_level
0	0.05	0.036667	0.18	8294	0.020595	0.064454
1	0.05	0.070000	0.17	9246	0.046237	0.104636
2	0.05	0.066667	0.16	10379	0.043568	0.100723
3	0.05	0.056667	0.15	11742	0.035677	0.088866
4	0.05	0.053333	0.14	13402	0.033092	0.084869

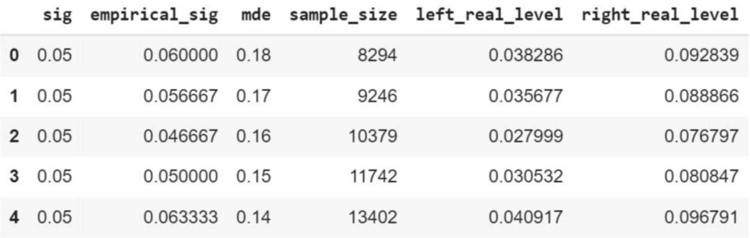
Рисунок 25 - A/A симуляция для z-test



	power	empirical_power	mde	sample_size	left_real_level	right_real_level
0	0.8	0.643333	0.18	8294	0.587630	0.695413
1	0.8	0.606667	0.17	9246	0.550375	0.660261
2	0.8	0.660000	0.16	10379	0.604674	0.711280
3	0.8	0.723333	0.15	11742	0.670130	0.770889
4	0.8	0.683333	0.14	13402	0.628659	0.733372

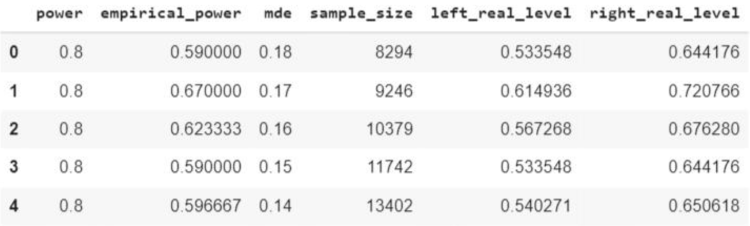
Рисунок 26 - A/B симуляция для z-test

The next step will be the bootstrap of the ratio metric. A/A and A/B simulations are shown in Figures 27 and 28. There are no noticeable improvements compared to the previous method.



	sig	empirical_sig	mde	sample_size	left_real_level	right_real_level
0	0.05	0.060000	0.18	8294	0.038286	0.092839
1	0.05	0.056667	0.17	9246	0.035677	0.088866
2	0.05	0.046667	0.16	10379	0.027999	0.076797
3	0.05	0.050000	0.15	11742	0.030532	0.080847
4	0.05	0.063333	0.14	13402	0.040917	0.096791

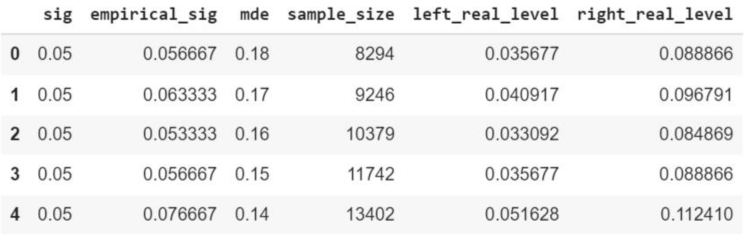
Рисунок 27 - A/A симуляция для теста Бутстрап



	power	empirical_power	mde	sample_size	left_real_level	right_real_level
0	0.8	0.590000	0.18	8294	0.533548	0.644176
1	0.8	0.670000	0.17	9246	0.614936	0.720766
2	0.8	0.623333	0.16	10379	0.567268	0.676280
3	0.8	0.590000	0.15	11742	0.533548	0.644176
4	0.8	0.596667	0.14	13402	0.540271	0.650618

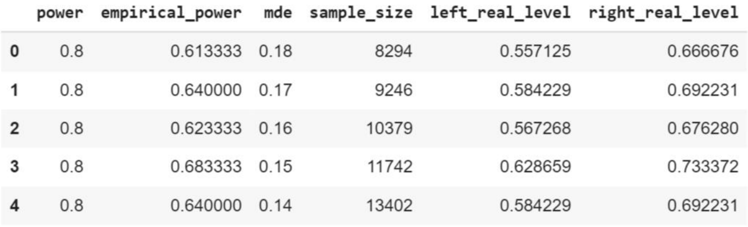
Рисунок 28 - A/B симуляция для теста Бутстрап

Figures 29 and 30 show the results for the delta method, which is the correct analogue of the proportional z-test. However, there are no noticeable improvements.



	sig	empirical_sig	mde	sample_size	left_real_level	right_real_level
0	0.05	0.056667	0.18	8294	0.035677	0.088866
1	0.05	0.063333	0.17	9246	0.040917	0.096791
2	0.05	0.053333	0.16	10379	0.033092	0.084869
3	0.05	0.056667	0.15	11742	0.035677	0.088866
4	0.05	0.076667	0.14	13402	0.051628	0.112410

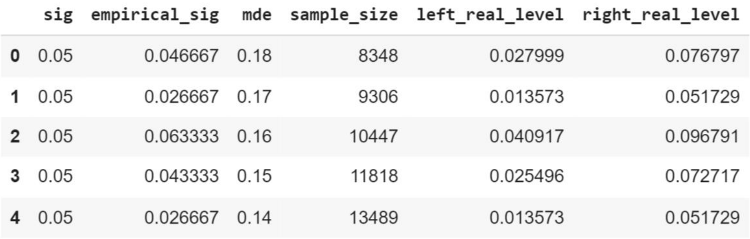
Рисунок 29 - A/A симуляция для дельта метода



	power	empirical_power	mde	sample_size	left_real_level	right_real_level
0	0.8	0.613333	0.18	8294	0.557125	0.666676
1	0.8	0.640000	0.17	9246	0.584229	0.692231
2	0.8	0.623333	0.16	10379	0.567268	0.676280
3	0.8	0.683333	0.15	11742	0.628659	0.733372
4	0.8	0.640000	0.14	13402	0.584229	0.692231

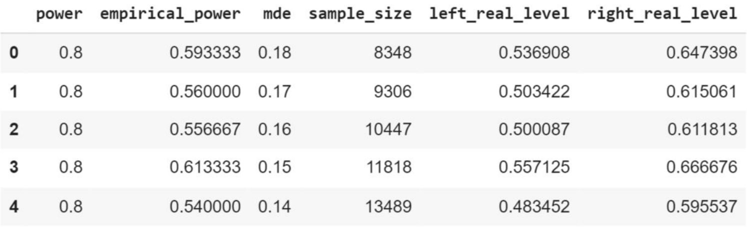
Рисунок 30 - A/B симуляция для дельта метода

Considering that we are dealing with a ratio metric, theoretically the best method in this case would be the test proposed in Keyu Nie's article “Dealing With Ratio Metrics in A/B Testing at the Presence of Intra-User Correlation and Segments". The results are shown in Figures 31 and 32. As you can see, there are no improvements, there are even minor impairments. It can be assumed that the reason for this is that the number of users who have more than one visit is very small, and we are unable to correctly capture the inter-user correlation. Figure 33 shows the breakdown of users by the number of visits.



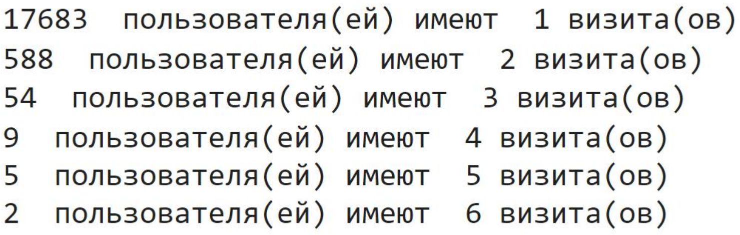
	sig	empirical_sig	mde	sample_size	left_real_level	right_real_level
0	0.05	0.046667	0.18	8348	0.027999	0.076797
1	0.05	0.026667	0.17	9306	0.013573	0.051729
2	0.05	0.063333	0.16	10447	0.040917	0.096791
3	0.05	0.043333	0.15	11818	0.025496	0.072717
4	0.05	0.026667	0.14	13489	0.013573	0.051729

Рисунок 31 - A/A симуляция для теста из статьи Keyu Nie



	power	empirical_power	mde	sample_size	left_real_level	right_real_level
0	0.8	0.593333	0.18	8348	0.536908	0.647398
1	0.8	0.560000	0.17	9306	0.503422	0.615061
2	0.8	0.556667	0.16	10447	0.500087	0.611813
3	0.8	0.613333	0.15	11818	0.557125	0.666676
4	0.8	0.540000	0.14	13489	0.483452	0.595537

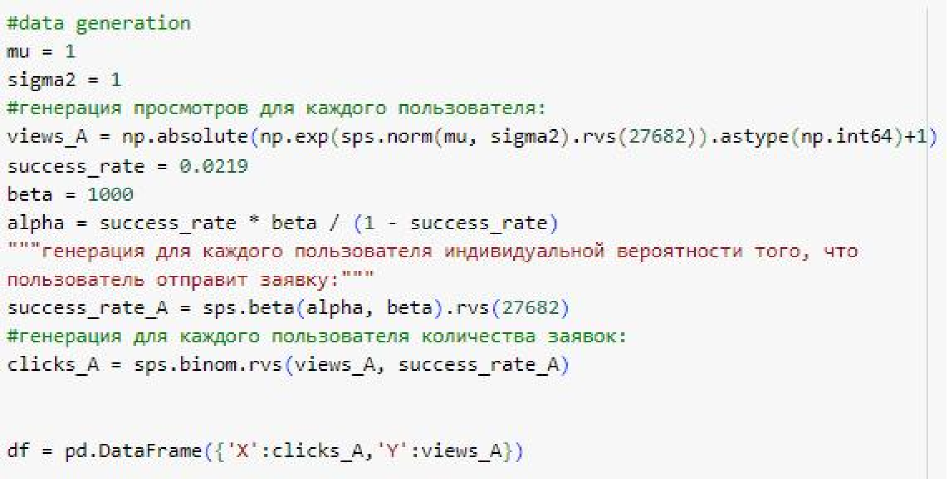
Рисунок 32 - A/B симуляция для теста из статьи Keyu Nie



```
17683  пользователя(ей) имеют  1 визита(ов)
588  пользователя(ей) имеют  2 визита(ов)
54  пользователя(ей) имеют  3 визита(ов)
9  пользователя(ей) имеют  4 визита(ов)
5  пользователя(ей) имеют  5 визита(ов)
2  пользователя(ей) имеют  6 визита(ов)
```

Рисунок 33 - разбивка пользователей по количеству визитов

To check the performance of this test, it is proposed to generate a dataset as shown in Figure 34. The number of users and conversion have the same values as in our dataset. The difference is that the number of user visits has a more diverse distribution, as shown in Figure 35. In Figure 36, we see the results of the simulation of the A/B test for this dataset. Now the empirical power has become close to one, this may be a confirmation of the assumption from the previous paragraph.



```
#data generation
mu = 1
sigma2 = 1
#генерация просмотров для каждого пользователя:
views_A = np.absolute(np.exp(sps.norm(mu, sigma2).rvs(27682)).astype(np.int64)+1)
success_rate = 0.0219
beta = 1000
alpha = success_rate * beta / (1 - success_rate)
"""генерация для каждого пользователя индивидуальной вероятности того, что
пользователь отправит заявку:"""
success_rate_A = sps.beta(alpha, beta).rvs(27682)
#генерация для каждого пользователя количества заявок:
clicks_A = sps.binom.rvs(views_A, success_rate_A)


df = pd.DataFrame({'X':clicks_A,'Y':views_A})
```

Рисунок 34 - генерация датасета

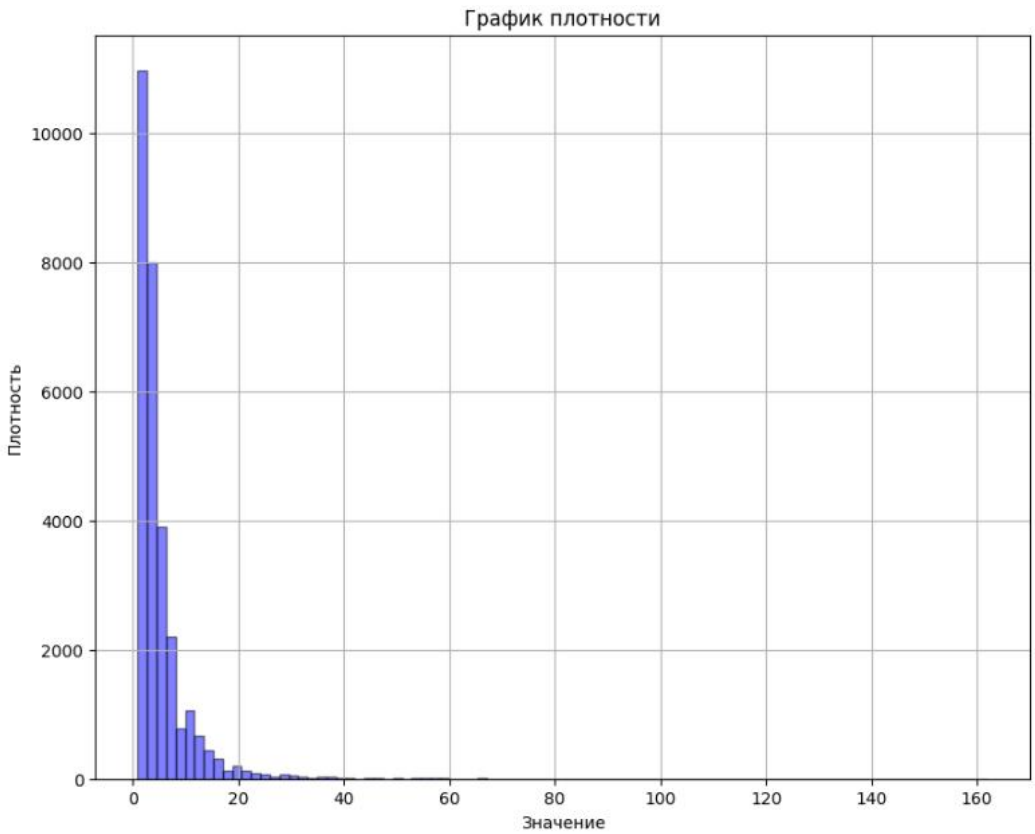
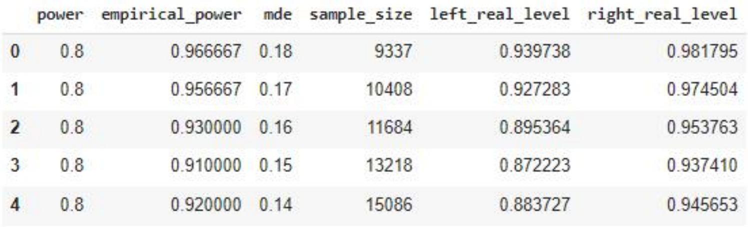


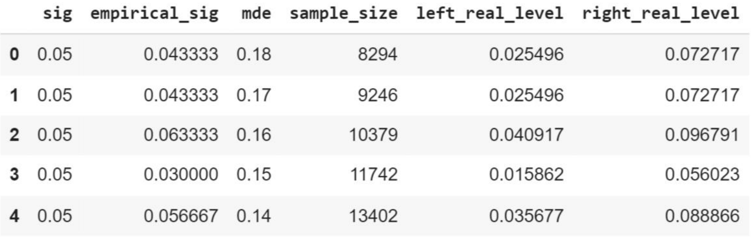
Рисунок 35 - распределение визитов в сгенерированном датасете



	power	empirical_power	mde	sample_size	left_real_level	right_real_level
0	0.8	0.966667	0.18	9337	0.939738	0.981795
1	0.8	0.956667	0.17	10408	0.927283	0.974504
2	0.8	0.930000	0.16	11684	0.895364	0.953763
3	0.8	0.910000	0.15	13218	0.872223	0.937410
4	0.8	0.920000	0.14	15086	0.883727	0.945653

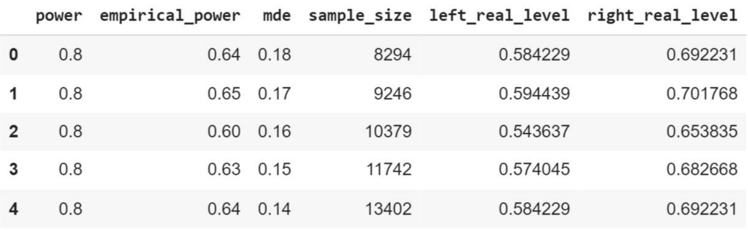
Рисунок 36 - A/B симуляция для сгенерированного датасета

The next approach, for which simulations were held on the original dataset, is linearization. The results are visible in Figures 37 and 38, and they also have no significant improvements.



	sig	empirical_sig	mde	sample_size	left_real_level	right_real_level
0	0.05	0.043333	0.18	8294	0.025496	0.072717
1	0.05	0.043333	0.17	9246	0.025496	0.072717
2	0.05	0.063333	0.16	10379	0.040917	0.096791
3	0.05	0.030000	0.15	11742	0.015862	0.056023
4	0.05	0.056667	0.14	13402	0.035677	0.088866

Рисунок 37 - A/A симуляция для линеаризации



	power	empirical_power	mde	sample_size	left_real_level	right_real_level
0	0.8	0.64	0.18	8294	0.584229	0.692231
1	0.8	0.65	0.17	9246	0.594439	0.701768
2	0.8	0.60	0.16	10379	0.543637	0.653835
3	0.8	0.63	0.15	11742	0.574045	0.682668
4	0.8	0.64	0.14	13402	0.584229	0.692231

Рисунок 38 - A/B симуляция для линеаризации

Simulations were also held for the multi-armed bandits method, the empirical powers are close to one for all MDEs. However, here we cannot control which MDE criterion should be considered statistically significant. The empirical significance level turns out to be approximately 0.5, since in the absence of a difference between the samples, the method chooses the winner randomly.

In view of all the above, the delta method was chosen for the A/B test due to its theoretical correctness, mde was decided to equate 0.15. As a result, the test showed that there was no statistically significant increase in conversion. Moreover, if our alternative hypothesis were the reverse of the current one, then the control version would be statistically significantly better than the test one, since the p-value is 0.953. To obtain this result, it took 11742 users for each sample.

Conclusion

Thus, according to the results of the conducted research:

- The analysis of the latest researches in the field of A/B testing of various types of metrics has been carried out. Conclusions were drawn about the advantages of certain methods
- The described methods are implemented at the software level.
- The most important metrics used in A/B testing of web forms have been explored.
- The analysis and preprocessing of the datasets used, highlighting their features. These features have been applied to achieve the best performance in A/B tests.
- A/B and A/A testing simulations are implemented to verify the operability of various methods at the software level.
- Full cycles of A/B tests have been performed for two datasets with different metrics.

Summary

A/B testing is a reliable tool for optimizing brand interaction with consumers in the modern world of digital technologies. In modern business, the most popular approach is scientifically based, which is what A/B testing is. Web forms are present on a variety of sites, and companies are interested in creating the most user-friendly options.

As a result of two A/B tests, one statistically significant result was obtained. The strengths and weaknesses of various methods were also highlighted. During the analysis of datasets, various features were highlighted that are important in the context of A/B testing.

Bibliographic list

1. Kohavi, R., Longobardi, R., & Sommerfield, D. (2016). Controlled experiments on the web: survey and practical guide. Data Mining and Knowledge Discovery

2. Lincoln Murphy: How to Use RICE to Prioritize Features and Product Enhancements

3. Nadim Nachar(2008). The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution

4. Rice, John A. (2006). Mathematical Statistics and Data Analysis (3rd ed.). Duxbury Advanced.(420-443, 477-485)

5. Siroker, Koomen. A/B Testing, The most powerful way to turn clicks into customers

6. Кокрен У. Методы выборочного исследования - М., 1976.

7. Huizhi Xie, Juliette Aurisset. Improving the Sensitivity of Online Controlled Experiments: Case Studies at Netflix.

8. Alex Deng, Ya Xu, Ron Kohavi. Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data

9. Simon Jackson. How Booking.com increases the power of online experiments with CUPED

10. Maryam Aziz. On Multi-Armed Bandit Designs for Dose-Finding Clinical Trials

11. Keyu Nie, Ted Tao Yuan, Yinfei Kong, Pauline Berry Burke. Dealing With Ratio Metrics in A/B Testing at the Presence of Intra-User Correlation and Segments

12. Roman Budylin, Alexey V.Drutsa , Ilya Katsev, Valeriya Tsoy.Consistent Transformation of Ratio Metrics for Efficient Online Controlled Experiments

13. Alex Deng, Ulf Knoblich.Applying the Delta Method in Metric Analytics: A Practical Guide with Novel Ideas

14. Ron Kohavi, Roger Longbotham, Dan Sommerfield, Randal M.Henne. Controlled experiments on the web: survey and practical guide

15. Ramesh Johari, Leonid Pekelis, Pete Koomen, David Walsh.Peeking at A/B tests

16. Peter O’Brien, Thomas R.Fleming. A Multiple Testing Procedure for Clinical Trials

17. Wenru Zhou, Miranda Kroehl, Maxene Meier, Alexander Kaizer. Approaches to analyzing binary data for large-scale A/B testing