Knowledge document

Change in A/B-test evaluation method

Innovation and keeping ahead of the curve are key drivers of Online Dialogue. We have decided to change our A/B-test evaluation method. We will briefly explain why and how we will operate from now on.

March 30, 2016

Online Dialogue

Change in A/B-test evaluation method

Innovation and keeping ahead of the curve are key drivers of Online Dialogue. In this document two main A/B-test evaluation methods are discussed. We have decided to change our evaluation method. Therefore we will briefly explain our reasons why and how we will operate from now on.

Description of test evaluation methods

Until recently Online Dialogue used – just as almost everyone in the market – frequentist statistics (with a t-test) to evaluate A/B-tests. A t-test checks whether the averages of two independent groups differ significantly from each other. The basic assumption of this test is that there is no difference in conversion rate between group A and B. This is the so-called null hypothesis. With a t-test you try to reject this hypothesis, since you want to prove that your test variation (B) outperforms the original (A). With a set significance level in advance of the test (usually 90 or 95 percent) you judge how unlikely the measured difference in the test between variation A and variation B is. If the result is very unlikely under the null hypothesis - say with a p-value of 0.02 - then you could safely state that the conversion rate of A is different from that of B. Only when a significant difference is measured, will the recommendation be to implement the variation.

With Bayesian statistics a test conclusion is less clear-cut. Based on a test result the exact probability is determined that the variation outperforms the current situation. Consequently, with Bayesian statistics a test result does not have a binary outcome (winner or no winner), but a chance between 0 and 100 percent. Depending on the risk you are willing to take, this could mean that with a chance of 80 percent or even 70 percent you decide to implement the variation. This seems like a smart decision, because the probability that the variation outperforms the original is higher than 50 percent. With this test evaluation method you will not only stick to implementing clear winners (of which you truly learn something), but also implement variations which will only indicatively increase revenue (but of which you cannot derive true behavioral insights from).

Change in A/B-test evaluation method

Why do we switch to a Bayesian evaluation method?

With a Bayesian evaluation method Online Dialogue expects to optimize the customer dialogue even better. This method ensures that test results and conclusions can be communicated without any statistical terminology. A Bayesian test evaluation gives a much simpler answer to the question whether variation B outperforms the current situation, namely with a chance.

This is far easier to understand and more relevant to the business question than 'how unlikely is the difference found, given that there is no difference' (the conclusion based on frequentist statistics). This is also one of the reasons why more and more A/B-test software packages are shifting towards Bayesian instead of frequentist statistics to evaluate test results.

A second important reason is that in practice we come across many cases where the statistics only weakly support that B is better than A (frequentist), but where implementing B would actually be a smart decision in order to make money (Bayesian). With a frequentist evaluation method it is only recommended to implement a test variation when a significant difference is found in an A/B-test. Consequently, test variations that only indicatively increase revenue will not be implemented.

With a Bayesian test evaluation the risk of implementing non-significant test variations is mapped out. Every test results in a risk assessment, where the expected extra revenue is evaluated against the risk that the variation actually underperforms. The positive effect is that more variations will be implemented, resulting in a higher revenue growth.

Change in A/B-test evaluation method

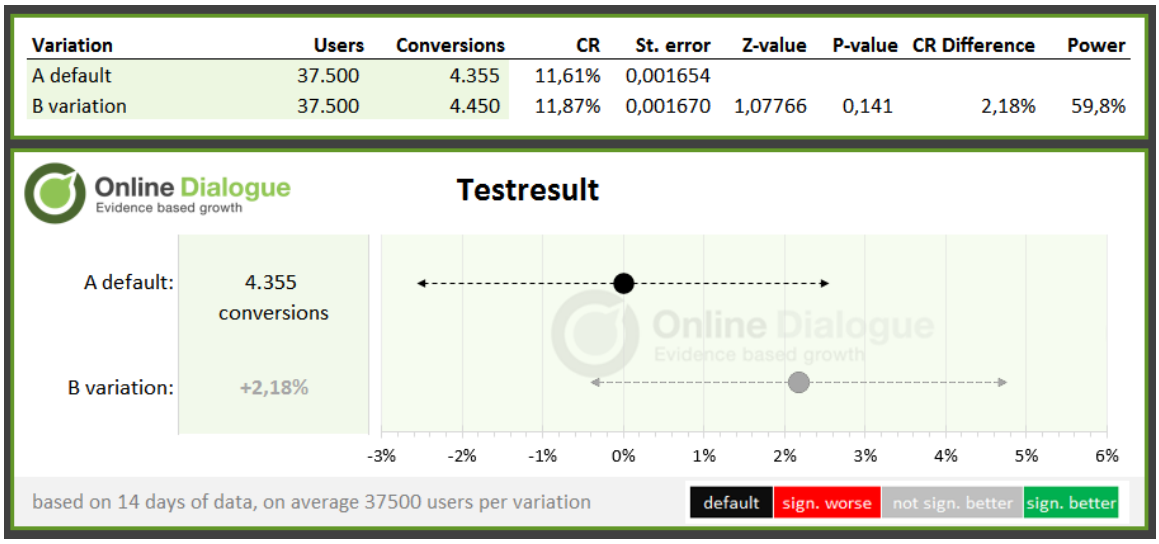
Case example

Assume that company X has carried out an A/B test. Each variation had 37.500 visitors and variation B had a measured conversion uplift of 2.18%.

Frequentist A/B-test evaluation

With a frequentist evaluation method the p-value determines whether or not the variation will be implemented.

Variation	Users	Conversions	CR	St. error	Z-value	P-value	CR Difference	Power
A default	37.500	4.355	11,61%	0,001654				
B variation	37.500	4.450	11,87%	0,001670	1,07766	0,141	2,18%	59,8%

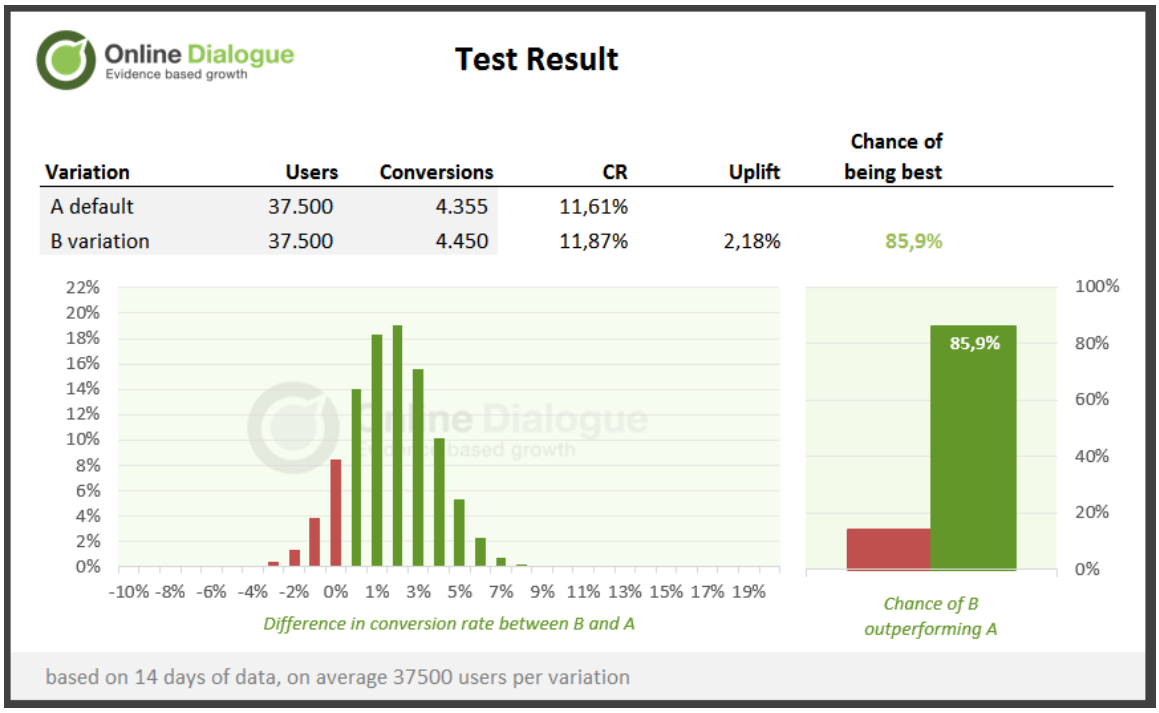


The p-value in this test was 0.141. Based on a significance level of 90%, the conclusion is that the conversion rate of variation B does not significantly differ from that of variation A, because the p-value (0.141) is higher than the cut-off value of 0.1. The difference found is simply not large enough to declare it a significant winner. The recommendation would be not to implement the variation and a different direction for future tests is probably chosen.

Change in A/B-test evaluation method

Bayesian A/B-test evaluations

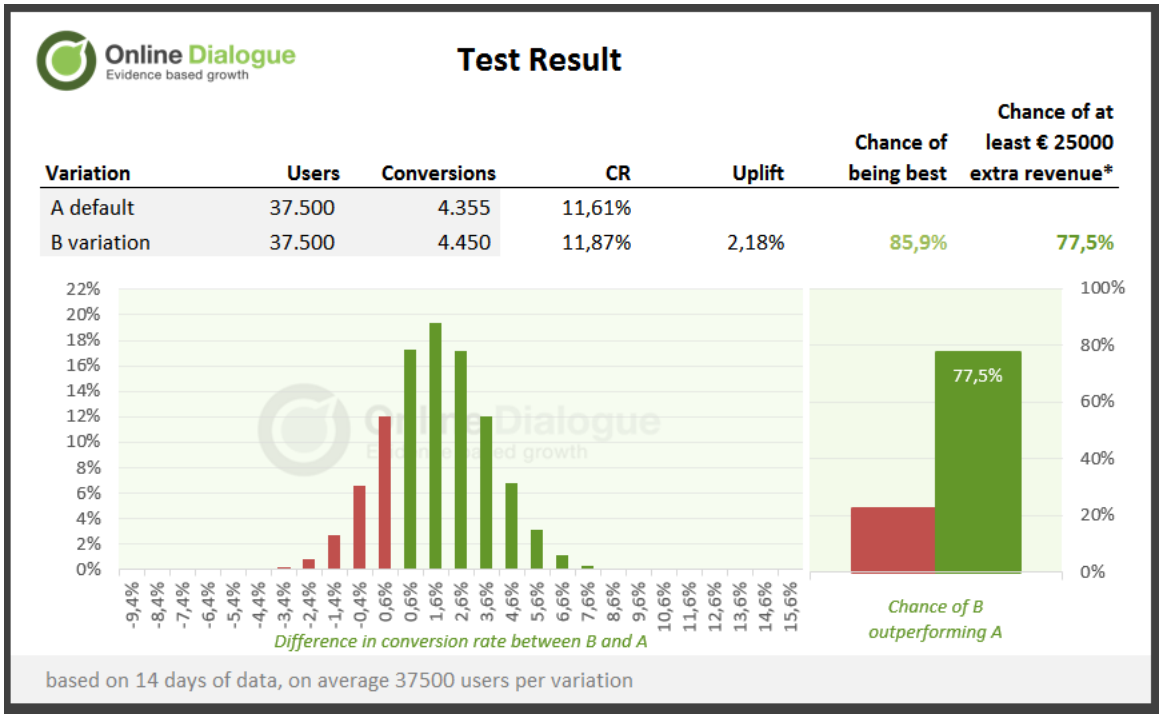
With a Bayesian test evaluation the chance that variation B outperforms the original is equal to 85.9%. The graph 'difference in conversion rate between B and A' shows that the difference in conversion rate is 85.9% of the time higher than 0%. The highest probability levels are in the range of +1% to +3%.



The question whether or not to implement the variation is a clear risk assessment based on the test outcome. Does a chance of 85.9% of an increase in revenue outweigh a risk of 14,1% to lose money? It could be argued that every test variation with a chance of 50% or higher should be implemented. However, costs of an optimization program and implementation costs should also be part of the equation for a fair assessment.

Change in A/B-test evaluation method

When the average order value and the minimum needed uplift in revenue of a test are known, it becomes also possible to calculate the chance of variation B bringing in that minimum revenue (for a good ROI). In this example the average order value is € 35 and the minimum uplift in revenue is € 25.000. This amount should be recouped in a six months period after implementation.



The chance of this scenario to occur is in this example 77.5% (this means a minimum uplift in conversion rate of 0.64%). These numbers give better input to make an informed decision whether or not to implement the variation. In this case there is a fair chance that the variation will be implemented. There is no significant difference in conversion rate found, but the test evaluation indicates that the test direction might be successful. In future tests this idea needs to be examined further.

Change in A/B-test evaluation method

Conclusion

With a Bayesian test evaluation Online Dialogue expects to optimize the customer dialogue even better. This evaluation method focusses A/B-test conclusions and course of actions on the business question at hand (will this variation make me more money?). It also makes communicating A/B test results more accessible and it will probably result in a higher revenue growth by really assessing the risk of implementing variations instead of a harsh cut-off between winner or no winner.

Questions?

For further explanation contact Annemarie Klaassen – Analytics & Optimization Expert at Online Dialogue

Telephone: +3130 7009 775

E-mail: annemarie@onlinedialogue.com